DOCUMENT RESUME

ED 186 223 SE 030 441

AUTHOR	Blank, A. A.; And Others
TITLE	Report of the Committee on Calculus. School Mathematics Study Group.
INSTITUTION	Stanford Univ., Calif. School Mathematics Study Group.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	63
NOTE	252p.; Contains light and broken type.
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	*Calculus; *Mathematics Curriculum; *Mathematics Instruction; Secondary Education; *Secondary School Mathematics; *Textbooks
IDENTIFIERS	*School Mathematics Study Group

ABSTRACT

Presented is the work of the calculus group of SMSG. The Advisory Board expressed a desire for a suitable high school calculus text. The calculus group surveyed available texts and, finding none which met advisory board requirements, outlined a suitable course. This outline was annotated for use by future writers. Some group members felt a more extended discussion of topics, suitable for use by students, was needed. A subgroup drafted this student version and in doing so changed the outline. The group could not reach agreement on these changes and therefore this document contains two reports, the second being the student version of the first four chapters of the calculus text. (MK)

****

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

****

NATIONAL SCIENCE FOUNDATION
COURSE CONTENT IMPROVEMENT
SECTION
OFFICIAL ARCHIVES
Do Not Remove From Office

**SCHOOL
MATHEMATICS
STUDY GROUP**

ED186223

REPORT OF THE COMMITTEE ON CALCULUS

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Mary L. Charles
of the NSF

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SE 030 441

© 1963 by The Board of Trustees of the Leland Stanford Junior University
All rights reserved
Printed in the United States of America

PREFACE

At the SMSG 1963 Summer Writing Session, a small group considered the question of a calculus text for high school use. The members of the group were:

A. A. Blank, New York University
R. J. Clark, St. Paul's School, Concord, New Hampshire
C. W. Leeds, Berkshire School, Sheffield, Massachusetts
N. L. Massey, Seattle Public Schools, Seattle, Washington
W. Stenberg, University of Minnesota, Minneapolis, Minnesota
R. D. Wagner, University of Wisconsin, Madison, Wisconsin

The SMSC Advisory Board had expressed a desire for a calculus text, suitable for high school use, which would presuppose the SMSG program through "Intermediate Mathematics" at least, and possibly through "Elementary Functions", and which would meet at least the level of the Advanced Placement Program.

The calculus group surveyed all available calculus texts to see if any of them satisfied the Advisory Board's requirements. Finding none, the group drew up an outline of a course which would, and then proceeded to annotate the outline.

The annotation consisted of discussions of the more important and more difficult topics. These discussions were intended for future writers, not for high school students. During the course of the session, however, some of the group felt that a more extended exposition of some of the topics should be prepared in a form suitable for students in order to indicate the intended style and pace of the student text.

Accordingly, a first draft of a student version of the first four chapters was prepared. The subgroup doing this felt that in order to make the material understandable to high school students, certain changes in the outline were necessary.

The group could not reach agreement in these changes, so two separate reports were prepared, the second one being the student version of the first four chapters.

PART A

Committee Exploring Calculus

Report

This report is divided into four sections, entitled The Need, The Survey, The Design, and The Samples. Although this makes for an orderly efficient report, it does not reflect the chronological progress of the committee. Given a broad mandate, the Committee Exploring Calculus (CEC) struggled with a number of ideas simultaneously and tried to find common areas of agreement. Although there were several points of view about the order of topics in the table of contents and the style of the samples, the conclusions in this report are supported by all members of the committee.

The Need

It is clear that each year more students will be capable of taking a full year of calculus in the 12th grade and will choose to do so. CEC does not assert any judgment about the wisdom and strength of this trend; sufficient for it to identify the fact and offer constructive measures. The committee does feel, however, that no calculus course should be less than a full year. There will be an increasing number of students studying calculus in the 12th grade and they will need suitable texts.

To fulfill this need, any text should be written in the style of SMSG and continue to exhibit sound mathematics. These two characteristics which reflect the underlying philosophy of the program have been an outstanding asset of the publications. The style can be summarized as self-teaching; i.e. requiring lucid, rather lengthy explanation, specifying all details and illustrated with numerous examples. The content consists of vital mathematics that reveals its inherent power and scope as well as its challenge. The problems must be fresh and interesting.

The text must cover in adequate depth the topics listed in the syllabus of the Advanced Placement Program. Since the APP and SMSG both seek to promulgate the mathematics that will best prepare a student for future work in the discipline, their syllabi will necessarily be very similar.

The text must present mathematics in such a way that each student will acquire an accurate, if limited, idea of analysis and appreciate the challenges of analytic methods. This is especially important in a course that may be a terminal course in mathematics for some. Unfortunately there are those few who take calculus and do not go to college. Also many choose other disciplines to the exclusion of mathematics when they enter college. Much has been done to alert citizens to the dichotomy between the scientific and liberal arts areas; the text should help to repair the breech. This is a need very peculiar to secondary school texts.

The Survey

CEC undertook a survey of existing texts to determine if there were any that sufficiently satisfied the need. The existence of such a text would preclude writing another.

At best, the evaluation of any textbook involves certain subjective difficulties, which the committee tried to minimize as each book was reviewed and discussed. No attempt was made to establish any check list with a weighting or point system.

The books examined included many very excellent texts and ranged from the highly sophisticated, very rigorous to the completely intuitive. The books generally fell into one or more of four categories.

Many were found to be too rigorous. Written at a high level of mathematical sophistication, such texts are not ideal for an introductory course. Others appeared to be simply too long. They include more topics than could be covered in a one year course. Generally these are books that were designed to be the text for programs of several semesters' duration.

A few books seemed to consist primarily of techniques with little supporting theory or background. By implication they suggest that the student needs only to categorize the problem and then repeat the exhibited solution. Finally some books were thought to be too brief or compact. While excellent books in a mathematical sense, they present severe pedagogical problems for any but the best teachers and certainly do not fulfull the need to be self-teaching.

Many of the books surveyed have been used as references in the study guide which CEC prepared to help teachers.

The unanimous opinion of the committee is that none of those surveyed (See Appendix A) satisfies the particular need. Therefore, CEC recommends that SMSG consider producing a text that does. To this end, the committee undertook to design a text and to produce samples.

The Design

The committee feels that a text designed to fulfill the need outlined earlier will have a number of special characteristics which are as follows:

A brief review. A uniform background is assumed. This means that little time is required to establish a framework for the course. Yet explicit statements of assumed foundations will enable classes of somewhat different background to begin with uniform symbolism and definitions. Also pedagogically, a review will be useful, perhaps necessary, to most secondary school teachers.

An assumption about analytic geometry. A student who has successfully completed SMSG work through the 11th grade has sufficient preparation to handle the concepts and problems. Certainly a full course in analytic geometry will strengthen a student's foundation. A review section on the general second-degree equation is contemplated and this can be expanded if an instructor desires.

Problem orientation. Insofar as possible, each major topic will be motivated by an actual problem for which there exists a mathematical model. Eventually, after a heuristic discussion, the empirical solution will be supported by theory and formalized since this approach must not be at the sacrifice of rigor.

Emphasis upon approximation. In elementary mathematics, methods appear to produce exact answers but the mature mathematician uses estimation a great deal. To instill an appreciation of approximations, especially with regard to the limit concept, is a major aim.

A formal treatment of limits. Following a rather lengthy discussion of the derivative on an intuitive basis, the formal ϵ, δ definition of a limit will be given. The ground will have been prepared, however, by introducing rather early the role that differences play by repeated use of such clauses as "x approaches a, which means that the difference $|x - a|$ can be made arbitrarily small." Following the formal definition of limit, some work will be given in ϵ, δ techniques. Then theorems on limits (of sums, etc.) will be displayed and thereafter used. Proofs of these theorems will either be in the text or in an appendix.

Separate development of the definite integral. Since a thorough understanding of this important topic is most important in subsequent mathematics, a slow careful development of the definite integral will be given. A completeness and a separation axiom about the real numbers will be given. The following separation axiom seems suitable. "If A and B are sets of numbers having the property that every member of A is less than or equal to every member of B, then there is a number s which separates A and B. $(x \in A \text{ and } y \in B \Rightarrow x \leq s \leq y)$ Although the antiderivative will be used sparingly earlier, the indefinite integral will not be introduced until after the Fundamental Theorem.

A teacher's commentary. As is customary with SMSG, the commentary will contain a rationale of what is done, suggestions to enhance the teaching, and answers together with explanations of all exercises.

The Samples

Sample sections of the proposed text (Appendix C) have been prepared so that the style, pace, and rigor can be judged. Although the samples have been rewritten to some extent, they are not to be considered final copy but rather a trial implementation of the objectives.

Since there was insufficient time for the small committee to write samples of all chapters, a table of contents (Appendix B) has been prepared to suggest the scope of the course. Believing that a Teacher's Commentary is an integral part of the program, CEC offers a small sample of this (Appendix D).

Appendix A

Textbooks Surveyed

Author(s)	Title	Date	Publisher
Adams, White	Analytic Geometry and Calculus	1962	Oxford
Agnew	Analytic Geometry and Calculus with Vectors	1962	McGraw-Hill
Andree	Introduction to Calculus	1962	McGraw-Hill
Apostol	Calculus	1962	Blaisdell
Bacon	Differential and Integral Calculus (2nd Ed.)	1955	McGraw-Hill
Begle	Introductory Calculus with Analytic Geometry	1954	Holt
Britton	Calculus	1956	Rinehart
Cogan, Horman, Thompson	Calculus of Functions of One Argument	1960	Prentice-Hall
Cooley	First Course in Calculus	1954	Wiley
Courant	Differential and Integral Calculus	1956	Interscience
Federer, Jonsson	Calculus	1961	Ronald
Fisher, Ziebur	Calculus and Analytic Geometry	1963	Prentice-Hall
Fobes, Smyth	Calculus and Analytic Geometry	1963	Prentice-Hall
Ford, Ford	Calculus	1963	McGraw-Hill
Franklin	Compact Calculus	1963	McGraw-Hill
Franklin	Differential and Integral Calculus	1953	McGraw-Hill
Goodman	Analytic Geometry and the Calculus	1963	MacMillan-Ginn
Granville, Smith, Longley	Elements of the Differential and Integral Calculus	1957	Ginn
Hart	Calculus	1955	Heath
Holmes	Calculus and Analytic Geometry	1950	McGraw-Hill
Johnson, Kiokemeister	Calculus with Analytic Geometry	1957	Allyn-Bacon
Kells	Analytic Geometry and Calculus	1950	Prentice-Hall
Kells	Calculus (2nd Ed.)	1949	Prentice-Hall
Kent	Differential and Integral Calculus	1960	Houghton-Mifflin
Kuratowski	Introduction to Calculus	1961	Addison-Wesley
Longley, Smith, Wilson	Analytic Geometry and Calculus	1960	Ginn
Love, Rainville	Differential and Integral Calculus (5th Ed.)	1954	MacMillan

Mask	An Introduction to Modern Calculus	1963	Holt-Rinehart-W.
Menger	Calculus	1955	Ginn
Merriman	Calculus	1954	Holt
Morrill	Calculus	1956	Van Nostrand
Morrey	University Calculus with Analytic Geometry	1962	Addison-Wesley
Peterson	Analytic Geometry and Calculus	1955	Harper
Phillips	Analytic Geometry and Calculus (2nd Ed.)	1946	Wiley
Protter, Morrey	Calculus with Analytic Geometry, A First Course	1963	Addison-Wesley
Randolph	Calculus and Analytic Geometry	1961	Wadsworth
Richmond	Calculus with Analytic Geometry	1959	Addison-Wesley
Sagan	Integral and Differential Calculus, An Intuitive Approach	1962	Wiley
Schock, Warshaw	Analytic Geometry and an Introduction to Calculus	1960	Prentice-Hall
Schwartz	Analytic Geometry and Calculus	1960	Holt-Rinehart W.
Sherwood, Taylor	Calculus (3rd Ed.)	1954	Prentice-Hall
Smail	Analytic Geometry and Calculus	1953	Appleton
Smail	Calculus	1949	Appleton
Smith, Salkover, Justice	Calculus	1958	Wiley
Sprague	Calculus	1952	Ronald
Thomas	Calculus and Analytic Geometry (3rd Ed.)	1960	Addison-Wesley
Thomas	Elements of Calculus and Analytic Geometry	1959	Addison-Wesley
Thurston	Calculus for Students of Engineering and the Exact Sciences	1963	Prentice-Hall
Townsend, Goodenough	Essentials of Calculus	1910	Holt
Wade	Calculus	1953	Ginn
Wylie	Calculus	1953	McGraw-Hill

Appendix B

Table of Contents

Chapter	Section	
1		Introduction
	1-1	A Problem Involving Area
	1-2	The Slope of a Curve
	1-3	Instantaneous Velocity
	1-4	The Limit Concept
2		Numbers, Functions and Graphs
	2-1	Introduction
	2-2	Real Numbers
	2-3	Functions
	2-4	Special Functions
	2-5	Second Degree Relations
3		The Derivative
	3-1	A Problem
	3-2	Numerical Computation of Slope
	3-3	The Slope Function for $y = x^2$
	3-4	The Velocity Function
	3-5	The Derivative
4		Limits
	4-1	Introduction
	4-2	Limit of a Sequence
	4-3	Limit of a Function
	4-4	Theorems on Limits
	4-5	Continuity
5		Differentiation
	5-1	Introduction
	5-2	Elementary Formulas
	5-3	The Chain Rule
	5-4	Implicit Differentiation
	5-5	Derivatives of Inverse Functions
	5-6	Circular Functions and their Derivatives
	5-7	Higher Order Derivatives

Chapter	Section	
6		Application of Differentiation
	6-1	Introduction
	6-2	Curve Tracing
	6-3	Maximum and Minimum Values on an Interval
	6-4	Optimum Value Problems
	6-5	Velocity and Acceleration
	6-6	Approximation and Differentials
7		The Law of the Mean and Its Consequences
	7-1	Introduction
	7-2	Rolle's Theorem
	7-3	The Law of the Mean
	7-4	Implications
8		The Definite Integral
	8-1	A Problem
	8-2	Measure of Closed Regions
	8-3	Approximation of Area by Summation
	8-4	Completeness Properties of Real Numbers
	8-5	The Definite Integral
	8-6	Properties of the Definite Integral
9		The Fundamental Theorem and Applications
	9-1	Introduction
	9-2	The Fundamental Theorem
	9-3	The Indefinite Integral
	9-4	Basic Formulas for Integration
	9-5	The Area Between Two Curves
	9-6	Volumes of Revolution
10		Logarithmic and Exponential Functions
	10-1	Introduction
	10-2	A Special Area Function
	10-3	The Natural Logarithmic Function
	10-4	Properties of Logarithmic Functions
	10-5	The Exponential Function
	10-6	Properties of Exponential Functions
	10-7	Differentiation and Integration

Chapter	Section	
11		Parametric Equations and Polar Coordinates
	11-1	Introduction
	11-2	Parametric Equations and their Derivatives
	11-3	Polar Coordinates
	11-4	Derivatives of Expressions in Polar Form
	11-5	Areas in Polar Coordinates
12		Further Applications of Integration
	12-1	Introduction
	12-2	Arc Length and Surface of Revolution
	12-3	Solids of Known Cross Section
	12-4	Volumes by Washers
	12-5	Cylindrical Shells
	12-6	Work
	12-7	Fluid Pressure
13		Techniques of Integration
	13-1	Introduction
	13-2	Substitution
	13-3	Powers of Trigonometric Functions
	13-4	Integration by Parts
	13-5	Inverse Trigonometric Functions
	13-6	Trigonometric Substitutions
	13-7	Partial Fractions
	13-8	Other Forms
14		Numerical Analysis
	14-1	Introduction
	14-2	Iteration Schemes, Newton's Method
	14-3	Approximate Integration
	14-4	Improper Integrals
	14-5	Approximation of Transcendentals by Polynomials
	14-6	Construction of Tables

Chapter	Section	
15		Applications
	15-1	Introduction
	15-2	Problems from Geometry
	15-3	Problems from Physical Sciences
	15-4	Problems from Other Sciences
	15-5	Theoretical Problems

PREFACE TO THE TEACHER

With the great number of calculus books now in print, what possible justification can there be for another? The justification is simple and direct; until now there has been no book written specifically for the student in the twelfth grade. This book is.

Many of the features special to this text arise directly from this writing group. The most notable feature is the care used in presenting a new topic. Particularly in a first course in analysis, time and patience must be given to the germination of the proper point of view. We seek to have a few topics understood in such depth that the power and subtlety can be appreciated and mastered sufficiently so that the colleges can build upon firm foundations rather than hasten along in order to display a number of empirical techniques.

A second feature of the text is motivation through problems. A desiccated theoretical discourse doesn't induce inquiring students to continue their study; they want to do things. Many major topics will be introduced by a problem. A pattern from problem to technique to extensions is used throughout the book.

A third feature is the treatment of the limit concept. From the beginning, the word difference is frequently used instead of the more usual words of "approach" or "gets close to." By using the form

$$\frac{f(x) - f(a)}{x - a},$$

the simultaneous consideration of the two differences $|f(x) - f(a)|$ and $|x - a|$ arises in an easy and natural way. The uses of numbers ϵ and δ are introduced separately at different places in the text so that confusion is minimized. Ultimately, after the derivative has been introduced on an intuitive level, a formal definition of limit is presented, and each instance of the derivative introduced earlier is proven formally. Having met and used both δ and ϵ in earlier situations, the student accepts and understands the more rigorous definition. Incidentally, two practices which experience has shown to promote confusion in an elementary course can be largely eliminated by this technique: (1) Δx does not "approach zero," a statement which combined with the idea of derivative at a leads to a false concept;

(2) x does not "go to infinity," a statement which promulgates a fantasy of motion that hinders a true understanding of limit.

A fourth feature is the emphasis upon the proper development of the definite integral. A completeness axiom for real numbers is given; the concept of upper and lower sums is developed and ultimately it is shown that $|\overline{S} - \underline{S}| < \epsilon$ whenever $0 < ||\Delta x|| < \delta$; finally the definite integral is defined as a number. Hopefully the "discovery" in Chapter 7 of the relationship between the integral and the derivative when the Fundamental Theorem is presented will be exciting.

The last features to be mentioned are the brevity of review and the sparcity of topics that "might" be included in a one-year course. It is assumed that anyone who starts a calculus course has studied and completed successfully the usual courses considered prerequisites. It is better to make this clear than to produce a lengthy review or introduce a compact presentation one chapter in length that will entice the improperly prepared. Also the presence of a number of extra topics, such as partial differentials and multiple integrals, militates against the original objective of a well-defined first course.

Hopefully this text will help you present calculus in such a way as to instill appreciation and promote understanding of some concepts and notions that have and do challenge some of the best minds of men.

Chapter 1

INTRODUCTION

Calculus has been, and is, a problem-oriented subject. The subject was developed as men sought methods for solving specific problems. As with all branches of mathematics, it is difficult, if not impossible, to fix upon a particular period in history as the moment of discovery or invention. It is expecially hard in the case of calculus since the problems that stimulated the creation of the subject had existed for thousands of years. The techniques were viewed pragmatically--if they produced answers they were considered sound. Subsequently, others analyzed the procedures in detail in order to know exactly what their predecessors were doing. It was through such analysis that the full power of the pioneer methods was realized.

In this first short chapter, we shall meet several classical problems whose solution led to the development of calculus. Our approach will be quite intuitive, but it will present the seed of the method which has become one of the most powerful branches of mathematics.

1-1. A Problem Involving Area

As you are aware, the area of a circle is found by using the formula

$$A = \pi r^2 .$$

We use decimal approximations for π in practical calculations. One commonly used is 3.14. Such an approximation was originally determined during a period of great mathematical activity in Greece (450 - 200 BC). Antiphon (420 BC) introduced a "method of exhaustion" that was developed by Eudoxus, Archimedes, and others. This process, which contains the seed of our modern calculus course, involves the use of a unit circle and regular polygons.

Let us start with one square inscribed and a second square circumscribed about a unit circle. See Figure 1-1a.

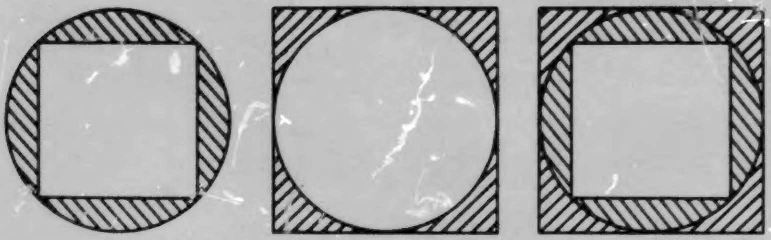


Figure 1-1a.

If we accept the notions that arise intuitively from our diagrams, we note that the area of the inscribed square is less than the area of the circle, the area of the circumscribed square is more than that of the circle, and the area of the circle must lie between the areas of the two squares. If we use squares, we have

$$2 < \text{area of circle} < 4.$$

Since $A = \pi r^2$ and since we have assumed $r = 1$, we immediately conclude

$$2 < \pi < 4$$

We haven't accomplished much so far, but we do have a logical argument for using a value, $\pi \approx 3$, in making a rough estimate! (The Book of Kings in the Bible contains just such an estimate.)

By increasing the number of sides of the regular polygons, we can make the interval containing π much smaller. For example, if we use hexagons (see Figure 1-1b) the shaded area in the last figure which represents the interval into which π is sandwiched is much smaller than the corresponding shaded area in the last diagram of Figure 1-1a.

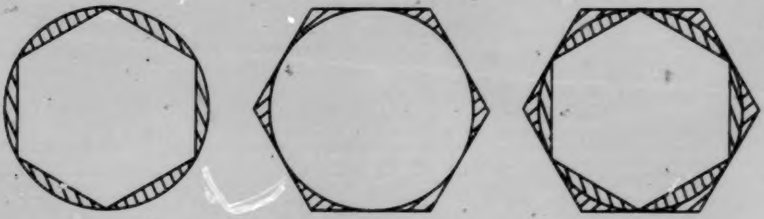


Figure 1-1b.

By actually computing the areas of the inscribed and circumscribed hexagons, we find

$$2.59 < \pi < 3.47$$

We can "squeeze" π between two numbers whose difference becomes smaller and smaller as we increase the number of sides in the regular polygons. For example, using polygons of 24 sides (see Figure 1-1c) we find that

$$3.105 < \pi < 3.172$$

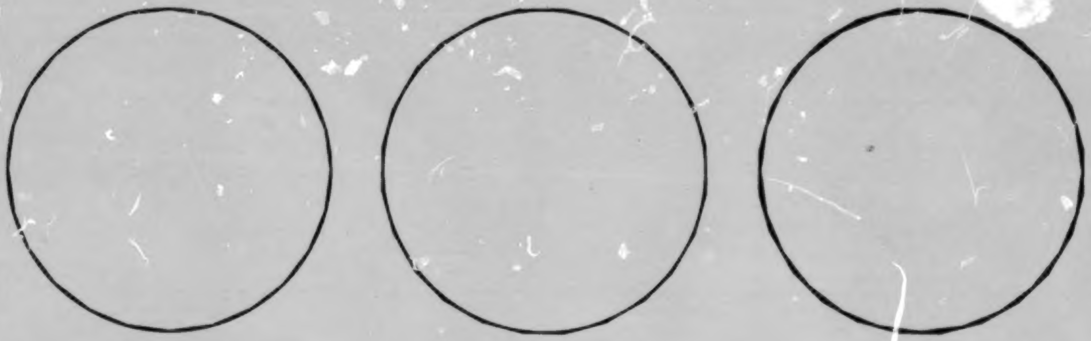


Figure 1-1c.

We can compute[*] the real number to which we assign the symbol π to any arbitrary accuracy by making the number of sides of the inscribed and circumscribed regular polygons sufficiently large.

In this illustration we have referred to the "method of exhaustion" which indicates how an area bounded by a curve can be approximated by areas whose boundaries are various polygons. Our study will reveal how this process, called the integral calculus, was refined and given sufficient power to make it one of the most important in the entire field of mathematics.

---

[*] The value of π, using less laborious methods, has recently been computed to 100,265 places, by Shanks and Wrench, July 29, 1961, on an IBM 7090. This, however, is still an approximation.

Exercises 1-1

1. Using Smith's "History of Mathematics" or Boyer's "The History of the Calculus" as a source, write a brief history of the "method of exhaustion" and the "quadrature of a circle."
2. Compute the lower and upper bounds for π found by using an inscribed and a circumscribed octagon.

1-2. Slope of a Curve

A second problem that served to nourish the invention of calculus was the task of defining what we mean by "slope of a curve" and then obtaining ways of computing the value for a wide assortment of examples. We know what the slope of a straight line is. Simply stated, it is the ratio

$$\frac{\text{rise}}{\text{run}}$$

has the same value for different segments of the same line. More exactly that the slope of a straight line passing through the two points $_0)$ and (x_1,y_1) is

$$\frac{y_0 - y_1}{x_0 - x_1}\,.$$

If we now turn to a curve, the situation changes dramatically. (See Figure 1-2a.) If we think vaguely in terms of

$$\frac{\text{rise}}{\text{run}}$$

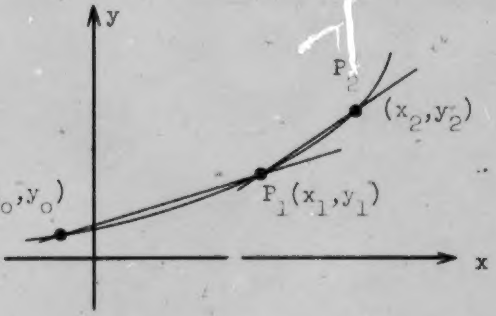


Figure 1-2a.

and choose three points on the curve, say P_0, P_1, and P_2, then a procedure similar to the straight line method yields the ratios

$$\frac{y_2 - y_1}{x_2 - x_1} \text{ and } \frac{y_1 - y_0}{x_1 - x_0} .$$

They are not necessarily equal, yet they might be. See Figure 1-2b.

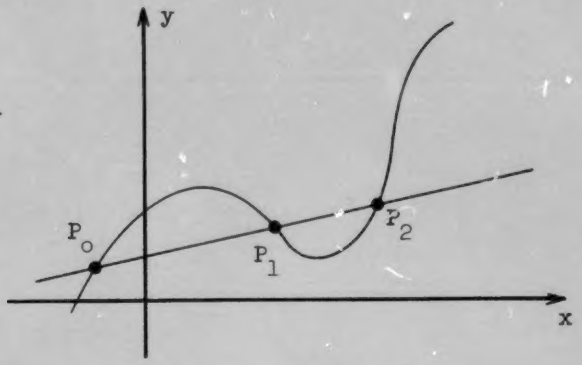


Figure 1-2b.

In this instance, however, the slope of the straight line is not what anyone would call the slope of the curve. The ratios

$$\frac{y_2 - y_1}{x_2 - x_1} \text{ and } \frac{y_1 - y_0}{x_1 - x_0}$$

represent the slopes of secant lines. What we seek is a measure of what we can vaguely call the "direction" of the curve.

If a boy is whirling a stone on the end of a string in a circular path and then releases the string, the stone flies off on a path which depends on the point of release. See Figure 1-2c.

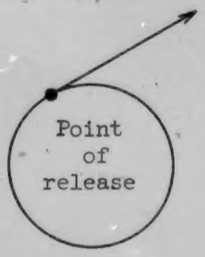


Figure 1-2c.

Somewhat the same problem is involved in safely launching a manned Mercury capsule. The rocket bearing the capsule is moving upward in a curved path. At the proper moment, the capsule is released and it continues along a path in the "direction" in which the powered rocket was moving. How we define and measure this "direction" of a curve is most pertinent.

What we want for our curves is something similar to what we had when we spoke of a tangent to a circle in our study of geometry. To draw such a line through a particular point of the circle, we constructed a line perpendicular to the radius through this point. See Figure 1-2d.

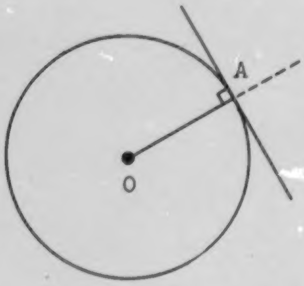


Figure 1-2d.

It helps to think of a tangent as being a straight line approximation of the circle in the neighborhood of point A. Since we are unable to draw radii and perpendiculars when we are dealing with other curves, we must develop a new approach to obtain a linear approximation. For example if we wish to draw a tangent, or linear approximation, to a curve through a point (x_1,y_1) then we consider a secant through this point and a second point close by on the same curve.

This second point can be on either side of (x_1,y_1) along the curve. See Figure 1-2e.

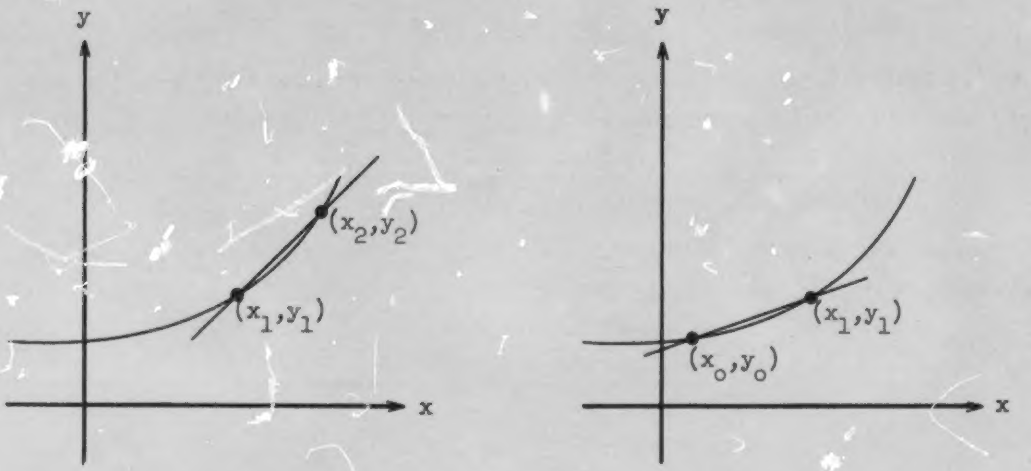


Figure 1-2e.

If we call these near-by points (x_2,y_2) and (x_0,y_0), and if we draw two secant lines, each through the given point and through one of the near-by points, then the slopes of these two lines are

$$\frac{y_2 - y_1}{x_2 - x_1} \text{ and } \frac{y_1 - y_0}{x_1 - x_0} .$$

Decreasing the distance between the two points (x_2,y_2) and (x_0,y_0), while keeping $x_0 < x_1 < x_2$, we can, in general, make the difference between the two slopes arbitrarily small. If we define our best linear approximation, or tangent, to the curve at point (x_1,y_1) as a line whose slope lies between the two slopes of such secant lines, we can find the slope of this tangent to any arbitrary exactness by taking points sufficiently close to (x_1,y_1).

The refinement and study of this second problem constitutes much of what is known as differential calculus or the calculus of derivatives. The discovery about 1670 by Leibnitz and Newton (at the same time and quite independently) of the relationship between the area problem and the tangent problem is generally called the starting moment of calculus.

Exercises 1-2

1. Consider the curve $y = x^2$. Find the slope of the secant line through each of the following pairs of points

   (a) (1,1) and (1.1,1.21)

   (b) (1,1) and (.9,.81)

   (c) (1,1) and (1.01,1.0201)

   (d) (1,1) and (.99,.9801)

2. Consider the curve $y = x^2$. What is the slope of a secant line through the two points $(a,f(a))$ and $(x,f(x))$.

1-3. Instantaneous Velocity

This third problem was the object of concern at the same period as the two we have considered previously, but we shall discuss it in a more modern setting. Consider a jet-liner that takes five hours to fly the 3000 miles from New York to San Francisco with Professor Begle aboard. The average velocity of the airplane is, therefore, 600 miles per hour, velocity being defined in the common manner,

$$\text{velocity} = \frac{\text{distance}}{\text{time interval}}$$

To Professor Begle, however, as the jet-liner rushes down the runway, the velocity of the jet-liner at the moment of take-off is more important. This critical velocity, 160 miles per hour, is not found by dividing the total distance by the total time. To determine if the jet-liner has reached the critical velocity, if it is going to fly, we must use a very small time interval near to the instant of take-off. The airplane takes off about 40 seconds after starting its run, but the distance on the runway divided by 40 doesn't yield 160 miles per hour, or its equivalent, 235 feet per second. We find the take-off velocity almost exactly by considering the distance covered in a very small time interval, one-half second for example, just before take-off. By choosing a time interval sufficiently small, we can obtain a velocity that we might arbitrarily label "instantaneous velocity."

Since it is exceedingly difficult to mark the exact moment that the jet-liner is beginning to fly, (do the wheels have to be completely without contact with the ground to have all the weight of the airplane borne by the wings?), it is useful to mark a position P_1 when the jet is definitely on the ground at time t_1, and then again mark the position P_2 when the jet is off the ground, at time t_2. An approximation of the instantaneous velocity can be made by finding the distance between the two positions, $P_2 - P_1$, and dividing this value by the time interval elapsed between t_1 and t_2, $t_2 - t_1$. By making the time difference sufficiently small, we make the difference quotient

$$\frac{P_1 - P_0}{t_1 - t_0}$$

approach, or be arbitrarily close to, the number which we choose to call instantaneous velocity.

This third problem is closely related to the one mentioned in Section 1-2. As we refine and develop the methods suggested in the description of these two problems, the differential calculus emerges and provides solutions not only to these, but also to a host of other problems.

Exercises 1-3

1. The following table contains the distances in feet a jet-liner covered from the starting point and the time in seconds required.

D	300	700	2500	4000	5200	5800	6000	6300	6500	
T	5	10	20	25	30	35	40	45	50	

(a) Find the average velocity during the first 10 seconds.

(b) Give a reasonable approximation to the instantaneous velocity at 10 seconds. What is the possible error of your approximation?

(c) If the jet-liner takes off after 40 seconds, what is its average velocity on the runway?

(d) From this table approximate the take-off velocity.

2. Suppose the distance, s, covered by a freely-falling body is determined by the equation

$$s = 32t^2$$

where t stands for time in seconds and s for distance in feet.

(a) Find the distance covered in 2 seconds.

(b) Find the distance covered in 2.1 seconds.

(c) Approximate the velocity of the freely falling body at the end of 2 seconds of free fall.

1-4. The Limit Concept

As the pioneers in the study of calculus obtained solutions to problems, they stimulated many others who, following in the steps of the pioneers, solved (circa 1700-1800) a wealth of assorted problems.

Subsequently, following the flood-tide of discovery, another generation became concerned (circa 1800-1900) about the theoretical aspects, and devoted its efforts to putting the subject on a firm or rigorous foundation. Foremost among those who made contributions were Gauss (1777-1855), Weierstrass (1815-1897), Cauchy (1789-1857), and Riemann (1826-1866). One of the difficult concepts to place upon a firm basis was the notion of a limit, and yet no aspect of the subject is more germane to a real comprehension of what is taking place.

In the discussion of the three problems presented earlier, we have been actually meeting the concept. The very word "exhaustion" implies a limit, even when used in other, non-mathematical, senses. The idea of "best" straight-line approximation to a curve suggests a limit; the term "instantaneous velocity" again brings up the concept. We should not conclude, however, that there are three fundamentally different limits involved. The problems are essentially applications of one central idea--a real number can be approximated within a controlled margin of error. This is the limit concept. Actually the concept is not a new one to anyone who has completed intermediate algebra since it was met when the sums of a geometric sequence are studied. Recall that a geometric sequence is defined as a sequence a, ar, ar^2, ar^3, ..., for example 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, ... Now if we consider the series

$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots,$ we speak of the sum as "approaching" 2. By controlling the number of terms n involved in forming a sum, the difference between the specific sum S_n and the "limit" 2 can be made as small as we please. If we let the Greek letter ϵ (epsilon) represent a small positive number, then we can summarize this discussion in mathematical notation, by writing

$$|S_n - 2| < \epsilon .$$

Employing the notion of differences to reach a useful definition of the word "limit" proves to be very fruitful as we shall see.

The origins of the derivative and the integral have been suggested in the discussion of the slope of a curve and the area of a circle respectively. The derivative is associated with variation; in general, it measures rate of change. Among the many interpretations of derivative we have velocity, acceleration, electrical current, heat flow, strain, density. The integral is associated with totality; it generally measures the end result or net effect of variation. It has interpretations such as the momentum acquired by a body affected by a force, electrical charge, energy, work, volume, mass. Later we shall see that derivative and integral are complimentary ideas and that the inverse relation between them can be exploited to great advantage. The point is not the universality of the two concepts above, but that there is a calculus, a system of reckoning, which enables us to solve important problems involving these ideas and to solve them simply and quickly. Just as science enriches mathematics by providing concrete models, mathematics enriches science by providing system and organization.

To develop this calculus we have made an intuitive beginning. The intuitive approach is useful and suggestive, but eventually we need to know just how far our methods work and when they are likely to fail. For this purpose we must frame our ideas precisely and reason about them logically, yet we shall not attempt to reduce the calculus to a complete deductive system. We shall try to label our omissions, however, so that you will be aware of the gaps to be filled if you undertake further study.

Exercises 1-4

1. Given the sequence $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ where $S_n = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{2^{n-1}}$, determine the number of terms that must be taken to make the difference between S_n and 2 less than

   (a) .01

   (b) .0001

   (c) 10^{-10}

2. Let $y = 2x^2 + x + 1$

   (a) If $x = 2.1$, find the difference $y - 11$.

   (b) If $x = 1.9$, find the difference $11 - y$.

   (c) If $x = 2.01$, find the difference $y - 11$.

   (d) If $x = 1.99$, find the difference $11 - y$.

   (e) If the difference $|11 - y|$ is to be made less than .01 what set of values can be assigned to x?

Chapter 2

NUMBERS, FUNCTIONS, and GRAPHS

2-1. Introduction

Before there can be communication between two people, there must be a language that is understood by both. Mathematics is a language which means that it has a vocabulary, grammar, and the other usual attributes. Compared with other languages, however, it is at once more precise and succinct. This precision, if fully achieved, assures us that no ambiguity can arise in the use of a word that is defined, nor in the application of a criterion in making decisions. Undoubtedly everyone who is embarking on this introductory course in calculus has studied algebra, plane geometry, trigonometry, and analytic geometry. These various branches, however, are not taught in a uniform way, nor should they be necessarily. In order to establish a common vocabulary, which is our immediate problem, we shall quickly review some familiar topics. The first of these will be real numbers, which are the basic objects to which our statements apply. On occasion, we shall use diagrams and graphs to provide an interpretation or an intuitive argument, yet the definitions and theorems must have meaning when attached only to real numbers and divorced from these pictures.

2-2. The Real Numbers

We are all familiar with many different sets of numbers from our past experience in mathematics. The numbers with which we shall be mainly concerned in elementary calculus are the real numbers. Subsets of the real numbers are: the set of natural numbers $\{1,2,3,\ldots\}$, the set of integers $\{\ldots,-3,-2,-1,0,1,2,3,\ldots\}$, the set of rational numbers including as elements $\frac{2}{5}$, $\frac{3}{4}$, and $\frac{0}{8}$, and the set of irrational numbers including as elements $\sqrt{2}$, $\sqrt{2+\sqrt{3}}$, and π. Unless specifically mentioned, we shall assume that the word "number" means "real number." The real numbers are characterized by three sets of properties: operation properties, order properties, and completeness properties.

We shall review the first two of these properties leaving the completeness to Chapter 8 where it will be important in the study of the definite integral.

The Operation Properties

For all real numbers a, b, c:

1.	Closure	$a + b$ is a real number ab is a real number
2.	Commutativity	$a + b = b + a$, $ab = ba$
3.	Associativity	$a + (b + c) = (a + b) + c$, [illegible]$(bc) = (ab)c$
4.	Identities	$a + 0 = a$, $a \cdot 1 = a$
5.	Inverses	$a + (-a) = 0$, $a \cdot (\frac{1}{a}) = 1$ where $a \neq 0$
6.	Distributivity	$a(b + c) = ab + ac$

These properties are only listed since they have been studied in previous courses.

Properties of Order

We recall that if a and b are real numbers, then exactly one of the following is true:

$$a < b, \quad a = b, \quad b < a$$

We define $a > b$, or $b < a$, to mean $a - b$ is a positive number. From this definition it follows that $a > 0$ is tantamount to saying that a is a positive number. The laws governing the usage of these inequality symbols are given below.

For all real numbers a, b, c:

5

1. If $a < b$ and $b < c$, then $a < c$
2. If $a < b$, then $a + c < b + c$

3. If $a < b$, then $ac < bc$ when $c > 0$

4. If $a < b$, then $ac > bc$ when $c < 0$

The expression $a < b < c$ means that $a < b$ and $b < c$ and we say that b is between a and c.

One of the basic assumptions in mathematics is that the real numbers can be placed in a one-to-one correspondence with the points on a line. Such a correspondence gives us a geometrical representation which serves to illustrate the order properties of the real numbers very clearly. If the numbers are assigned to the line in the usual way, positive to the right and the line horizontal, then $a > b$ is equivalent to stating that a is to the right of b on the number line. Thus not only is $4 > 3$ but we also say 4 is to the right of 3 on the number line. See Figure 2-2a.

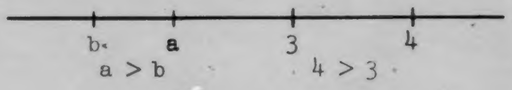


Figure 2-2a

If $a \neq b$, then between points a and b on the number line there is an interval. We need to develop some language and symbolism for talking about intervals. It is customary to use (a,b), where $a < b$, to mean "the set of all numbers between a and b." It will always be clear from the context whether the notation (a,b) represents an open interval or an ordered pair. Another way of denoting the same interval is the expression

$$\{x \mid a < x < b\} .$$

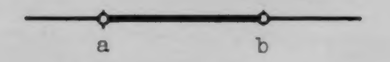


Figure 2-2b

This last expression, $a < x < b$, simply states the conditions necessary for x to be in the interval (a,b). Intervals of this form are called open intervals since the end points are not included. If we include the end points, then we write

$$[a,b] \quad \text{or} \quad a \leq x \leq b$$

and call this a closed interval. See Figure 2-2c.

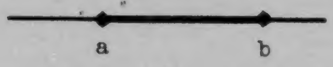


Figure 2-2c

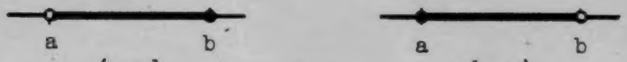
If we include one end point of an interval and not the other, we write

$$(a,b] \quad \text{or} \quad a < x \leq b$$

and

$$[a,b) \quad \text{or} \quad a \leq x < b \; .$$

These are called half-open intervals. See Figure 2-2d.



(a,b]

[a,b)

Figure 2-2d

Associated with any interval is the concept of length or distance between the points a and b. We shall let the distance, D, between a and b be defined as

$$|b - a| \; .$$

We know $|a - b| = |b - a|$, since for all numbers a

$$|a| = a \quad \text{if} \quad a \geq 0$$

and

$$|a| = -a \quad \text{if} \quad a < 0 \; .$$

The non-negative number thus defined is the length of the line segment from a to b.

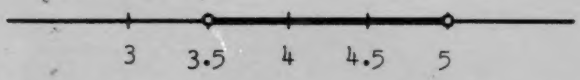


Figure 2-2f

A closed interval, such as $[1,13]$ in Figure 2-2g

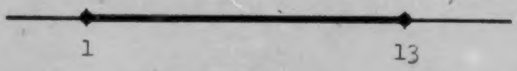


Figure 2-2g

can be expressed as either

$$\{x \,\|\, |x - 7| \leq 6\}$$

or $$\{x \,|\, 1 \leq x \leq 13\}$$

From the above examples we see that an inequality such as $|x - 5| < 3$ is equivalent to

$$-3 < x - 5 < 3$$

or $$2 < x < 8 \, .$$

In general, the set determined by the inequality

$$\{x \,\|\, |x - a| < \delta\},$$

where δ is any positive number, determines an open interval of length 2δ and with midpoint a as follows:

$$|x - a| < \delta$$

$$\Longleftrightarrow \quad -\delta < x - a < \delta$$

$$\Longleftrightarrow \quad a - \delta < x < a + \delta \, .$$

This interval is shown geometrically in Figure 2-2h.

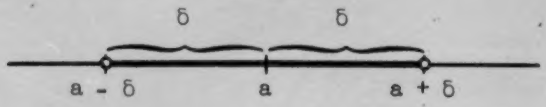


Figure 2-2h

Thus the number x is in the interval $|x - a| < \delta$ if and only if $a - \delta < x < a + \delta$. This is the open interval $(a - \delta, \ a + \delta)$.

Exercises

1. Use the definition of $a > b$ to prove that if $a > b$ and $b > c$, then $a > c$.

2. Use the definition of $a > b$ to prove that if $a > b$ and $c < 0$, then $ac < bc$.

3. If $a < b$, which of the following statements are true? false? indeterminate?

   (a) $7a < 7b$

   (b) $\frac{2}{3}a > \frac{2}{3}b$

   (c) $a + 2 < b + 3$

   (d) $a^2 < b^2$

   (e) $-4a < -4b$

   (f) $\sqrt{|a|} < \sqrt{|b|}$

4. Find the values of x for which the following inequalities are true and show these values on the number line.

   (a) $|x - 3| < 12$

   (b) $|x + 3| < 12$

   (c) $|x^2 + 2x - 6| < \frac{1}{2}$

   (d) $0 < |x - 2| < 5$

   (e) $|x| \leq 4$

   (f) $|x| \geq 4$

   (g) $|x + 3| \leq 0$

5. Show that $|a + b| \leq |a| + |b|$, for any a, b.

6. Show that $|a - b| \geq |a| - |b|$, for any a, b.

2-3. Functions

The effort of the scientist to understand our environment and that of the engineer to control it lead repeatedly to the attempt to determine some quantity unambiguously in terms of others. For example, an astronautical engineer who calculates the position of an orbiting satellite may fix its location if he knows the time elapsed since the launching rockets cut off, the point where cut-off occurred and the speed and direction of motion at the instant of cut-off. To the engineer it is imperative to know that this information is sufficient to determine the position of the satellite; in other words, that there is a functional dependence of position on the other data. Examples of this kind could be multiplied endlessly, but it is clear enough from this typical instance that the elementary concept of functional dependence permeates the body of scientific thought.

A view of the idea of functional dependence may be useful to jog our memories. We say, a datum y is functionally dependent upon data $x_1, x_2, \ldots, x_n$ if each assignment of specific values to the data x_i determines y uniquely. The correspondence between y and the x_i is defined as a function and we write

$$y = f(x_1, x_2, \ldots, x_n)$$

or, equivalently,

$$f: (x_1, \ldots, x_n) \longrightarrow y$$

to indicate the functional dependence. Both expressions may be read "f is the function which maps $(x_1, \ldots, x_n)$ onto y." Often y is referred to as the image of $(x_1, \ldots, x_n)$. For example, the area A of a triangle is functionally dependent upon the altitude h and the base b:

$$A = \frac{1}{2}bh.$$

An instructive example is the record of atmospheric pressure as a function of time plotted by a barograph at a fixed weather station. The pressure is functionally dependent upon the time since at any specific time in the historical record the pressure is uniquely determined. Clearly, functional dependence does not necessarily imply causal relation as in the satellite problem; the pressure can hardly be said to be caused by the time. Furthermore, functional dependence does not imply any law or rule like that determining the area of the triangle. There is no known rule for specifying the pressure at a given time apart from the historical record; we neither know what the atmospheric pressure was five hundred years ago nor precisely what it will be next week.

In this course, we shall treat only functions of the form $x \longrightarrow y$ where x and y are real numbers. It is not necessary that a functional dependence be defined for all real values of x. For example, the function

$$f: x \longrightarrow \sqrt{1 - x^2}$$

is defined only for x satisfying $-1 \leq x \leq 1$. The set of values of x for which the function is defined is called the domain of the function. The image of x is denoted by $f(x)$ so that we write in this particular instance

$$f(x) = \sqrt{1 - x^2}$$

and, similarly, for specific values of x we may write

$$f(1) = 0$$

$$f(-\tfrac{3}{5}) = \tfrac{4}{5}$$

$$f(\tfrac{1}{2} + h) = \sqrt{1 - (\tfrac{1}{2} + h)^2} = \tfrac{1}{2}\sqrt{3 - 4h - 4h^2}$$

$$f(0) = 1, \text{ etc.}$$

The set of images $f(x)$ for x in the domain of definition of f is called the range of the function. For the function $f\colon\ x \longrightarrow \sqrt{1 - x^2}$ the range consists of all values y satisfying $0 \leq y \leq 1$.

It is usually convenient to think of a function in terms of its graph, that is, the set of points (x,y) such that $y = f(x)$. See Figure 2-3a.

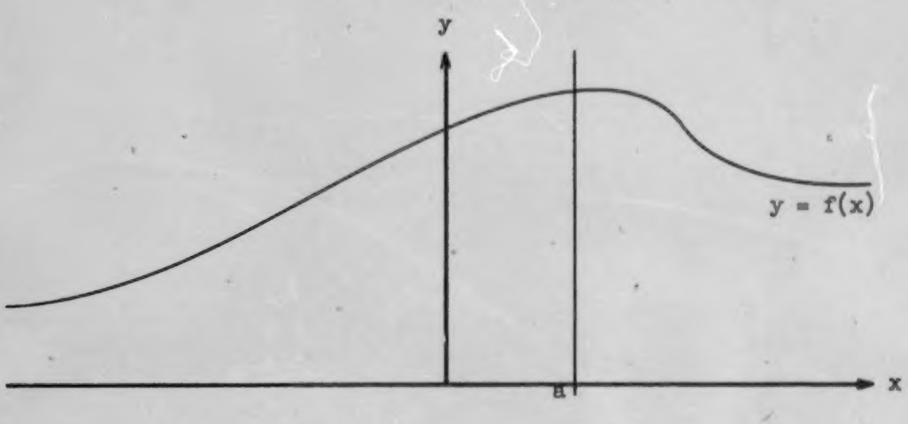


Figure 2-3a

The property of a function that each value of x in the domain determines just one value of y is reflected geometrically in the fact that a vertical line, $x = a$, intersects the graph in no more than one point. In other words, the graph of a function is a set of points such that no two points can have the same x-coordinate. This consideration leads to a definition of a real function as a set of ordered pairs of real numbers such that no two pairs have the same first member. We shall not concern ourselves with this definition except to note that the choice of a first member or element of the domain of the function uniquely fixes the second member, or element of the range and therefore a functional dependence is obtained under the conditions of the definition.

The first definition of a function, namely the one that defines a function as a correspondence, lends itself to interpretations in the abstract. We can draw diagrams and refer to elements in set A associated with, or mapped onto, elements in set B. See Figure 2-3b.

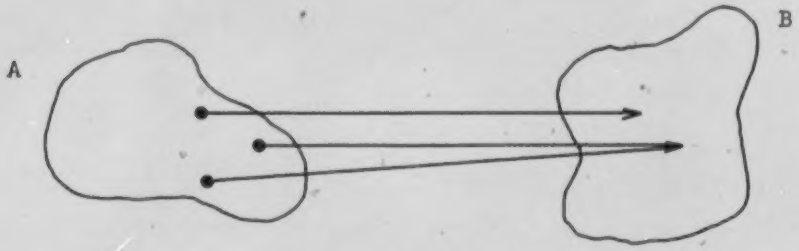


Figure 2-3b

Also we can speak of two number lines, one containing the domain of the function and the other containing the range. See Figure 2-3c.

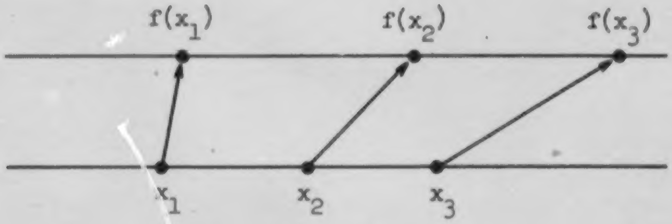


Figure 2-3c

Members of the domain x_1, x_2, x_3 ... are associated with their images $f(x_1)$, $f(x_2)$, $f(x_3)$ ... according to some rule, which is usually expressed as an equation. Each of the elements $f(x_1)$, $f(x_2)$, $f(x_3)$ ... is a member of the range.

We must bear in mind that an elliptical statement such as "the function $y = f(x)$" is an abbreviation for a statement such as "the function f such that each element, x, in the domain is associated with its image, y, in the range by the equation $y = f(x)$.

One of the more interesting properties of functions allows us to apply the four fundamental algebraic operations and create new functions. Suppose f and g are functions. Then we may have

cf where $y = cf(x)$,
$f + g$ where $y = f(x) + g(x)$,
$f - g$ where $y = f(x) - g(x)$,
and f/g where $y = f(x)/g(x)$.

Be careful, however, not to assume that the domain and range may be blithely combined in the same fashion. Obviously the domain of f/g contains only those numbers for which $g(x) \neq 0$.

We also speak about an inverse function in a manner similar to the way we speak about an inverse with numbers. Suppose we have a function f such that $f(x) = 2x + 1$. Thus we have, for example,

$2 \longrightarrow 5$,
$-1 \longrightarrow -1$,
$0 \longrightarrow 1$,
and $4 \longrightarrow 9$

The inverse function will map each element in the range of f, that is 5, -1, 1, 9 ..., back onto the associated elements in the domain. Thus the inverse function, which we call g, gives us

$5 \longrightarrow 2$,
$-1 \longrightarrow -1$,
$1 \longrightarrow 0$,
and $9 \longrightarrow 4$.

In each case, $g(x) = \frac{x - 1}{2}$. Every function does not have an inverse. For example, if $f(x) = x^2 + 1$, then

$1 \longrightarrow 2$,
and $-1 \longrightarrow 2$.

By definition, we cannot have a function that maps 2 onto both -1 and 1. Any time we have a function which sets up a one-to-one correspondence between the elements of the domain and the elements of the range, this function has an inverse.

An operation on functions that has no counterpart in the algebra of numbers is composition. If a function f associates a number x_1 with another number x_2, $(x_2 = f(x_1))$ and then a second function g associates x_2 with x_3, $(x_3 = g(x_2))$, then the composite function gf associates x_1 with x_3, $(x_3 = g(f(x_1)))$. It is important to bear in mind that gf does not necessarily equal fg. Note that when we write "gf" we mean f is to be applied before g and g applied to f(x). Specifically, let us consider f: $x \longrightarrow 3x - 1$, g: $x \longrightarrow 2x^2$, and the number 4. Under gf, we have first $f(4) = 11$ and then $g(11) = f(g(4)) = 242$. Under fg, we have first $g(4) = 32$ and then $f(32) = f(g(4)) = 95$. Since it is an apt expression, frequently a composite function is called "a function of a function."

If we think of a function as a machine with numbers of the domain as the input and numbers of the range as the output, we see that we can arrange two machines in tandem. See Figure 2-3d. The total

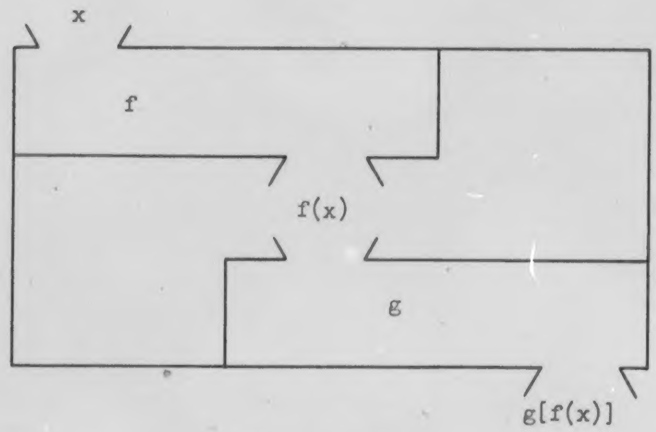


Figure 2-3d.

machine will perform the work of function gf. We may also represent a composite function by a mapping as follows:

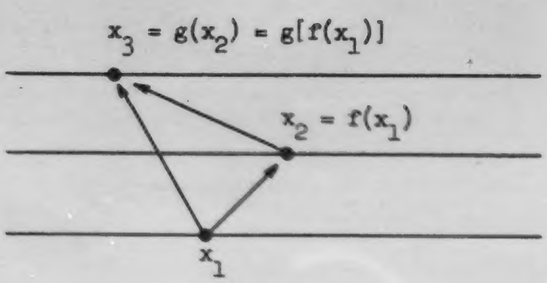


Figure 2-3e.

The most illuminating way of representing a specific function is through a graph. A graph of a function f is a set of points such that for each point (x,y) we have y = f(x). Usually our coordinate system will be the familiar rectangular or Cartesian coordinates, in which the horizontal axis is a number line containing the domain and the vertical line the range. We note now, although their use will be postponed, that polar coordinates are a second system that we can and will use.

2-4. Special Functions

In this section we shall consider an assortment of functions that occur often in a first-course in calculus. They merit some attention.

The simplest of functions is the constant function

$$f: x \longrightarrow c .$$

The domain consists of the entire set of real numbers; the range, of the single element, f(x) = c. Since all elements of the domain are associated with the one element of the range, there can be no inverse. The graph is a straight line parallel to the x axis .

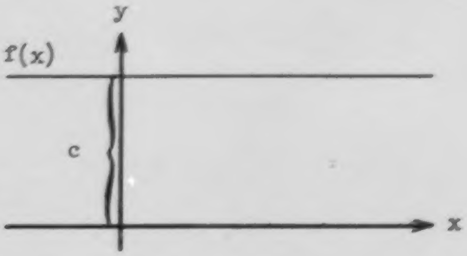


Figure 2-4a

41

A function f is a linear function if, for m, b real numbers, $m \neq 0$,

$$f : x \longrightarrow mx + b \, .$$

For any two pairs, $(x_1, f(x_1))$, $(x_2, f(x_2))$ that belong to the function, the ratio

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \qquad x_2 \neq x_1$$

is a constant. The constant is called the slope of the straight line we obtain whenever we graph a linear function.

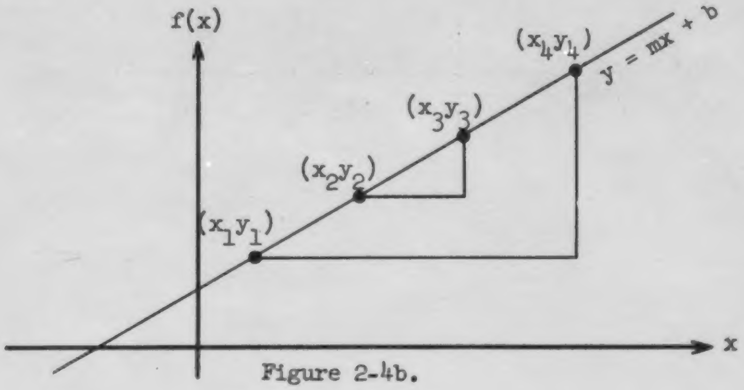


Figure 2-4b.

From similar triangles, (see Figure 2-4b) we have

$$\frac{y_4 - y_1}{x_4 - x_1} = \frac{y_3 - y_2}{x_3 - x_2} \qquad x_4 \neq x_1, \text{ and } x_3 \neq x_2$$

which supports our previous declaration.

Since we shall be concerned with the change in value of a function so frequently, we define the symbol "Δ" to indicate change. Thus Δx, "change in x" or simply "delta x", represents a change in the value of x. Corresponding to a change in the value of x will be a change in the value of the function determined by the difference

$$f(x + \Delta x) - f(x).$$

If $y = f(x)$, $\qquad y + \Delta y = f(x + \Delta x)$

and $\qquad \Delta y = f(x + \Delta x) - f(x) \, .$

Using this notation we can write

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x} . \qquad \Delta x \neq 0$$

If we are concerned with a linear function (see Figure 2-3a), we have

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x} = \text{constant } (m). \quad \Delta x \neq 0$$

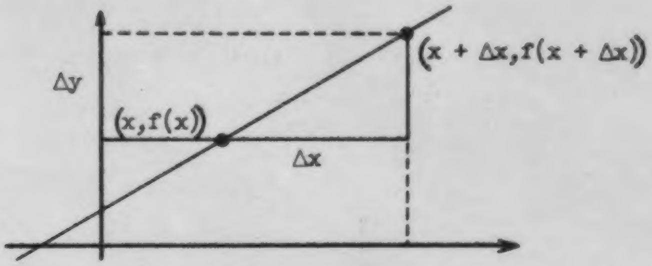


Figure 2-4c

Note that we are not restricting Δx to positive values; we can have a positive or negative change.

Example. Find the point common to the graphs of the two functions,

$$f: \ x \longrightarrow 3x + 1$$
$$g: \ x \longrightarrow -\frac{2}{3}x + 6;$$

the angle of intersection of the graphs; and the area enclosed by the graphs and the x-axis.

Solution. See Figure 2-4d. First we graph the two functions. The points $(0,1)$ and $(-2,-5)$ determine one line; $(0,6)$ and $(6,2)$ the other.

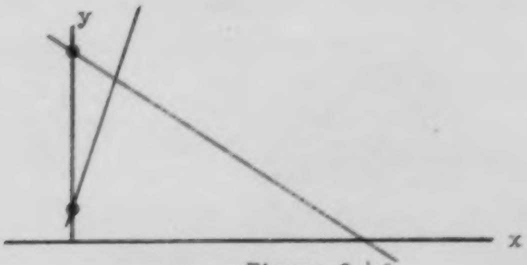


Figure 2-4d.

Solving simultaneously we have

$$\begin{array}{rl} 2y = & 6x + 2 \\ 9y = & -6x + 54 \\ \hline 11y = & 56 \\ y = & \frac{56}{11} \end{array} \qquad\qquad \begin{array}{rl} y = & 3x + 1 \\ -y = & \frac{2}{3}x - 6 \\ \hline 0 = & \frac{11}{3}x - 5 \\ \frac{15}{11} = & x . \end{array}$$

The point of intersection is $(\frac{15}{11}, \frac{56}{11})$. The angle of intersection, θ, is determined by the formula

$$\tan\theta = \frac{m_2 - m_1}{1 + m_2 m_1}, \qquad m_1 \neq \frac{-1}{m_2}$$

where m_2 and m_1 are the slopes of the lines.

$$\tan\theta = \frac{\frac{2}{6} - 3}{1 - 3} = \frac{11}{6} .$$

$$\theta \approx 61^\circ\ 23' .$$

The area is found through the formula $\frac{1}{2}bh$.

$$b = |9 + \tfrac{1}{3}| = 9\tfrac{1}{3} .$$

$$\frac{1}{2}bh = \frac{1}{2}(9\tfrac{1}{3})(\tfrac{56}{11}) = \frac{784}{33} \approx 23.8 .$$

Two more functions that have special interest for calculus students are the absolute value function and the step function. The absolute value function

$$f: x \longrightarrow |x| \qquad \text{for all } x \in R$$

has as its domain all real numbers x, but the range is all positive numbers $0 \le f(x)$. Its graph is shown in Figure 2-4e.

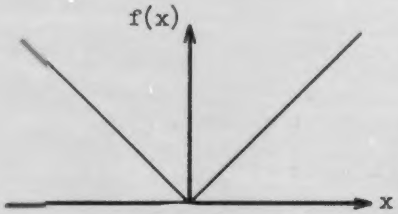


Figure 2-4e

The greatest integer function is defined by $f(x) = [x]$ where $[x]$ denotes the greatest integer n such that $n \leq x$. This is a function quite different from those with which we are more familiar. As examples of $[x]$ we see that $f(1) = [1] = 1$, $f(\frac{3}{2}) = [\frac{3}{2}] = 1$, $f(\sqrt{2}) = [\sqrt{2}] = 1$, $f(\pi) = [\pi] = 3$, $f(3) = [3] = 3$, and $f(3.1) = [3.1] = 3$. See Figure 2-4f.

In this case the domain is the set of all real numbers x, and the range is only the integers.

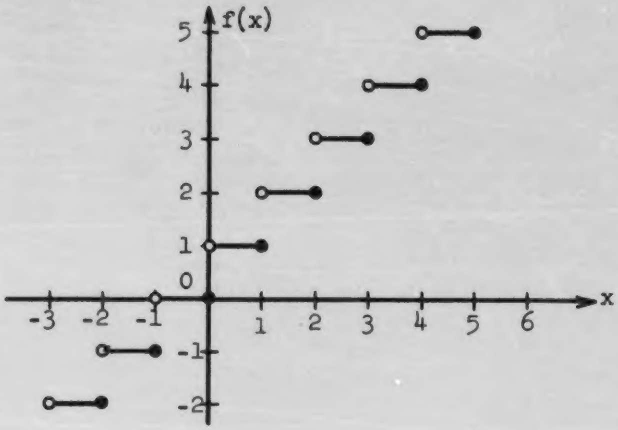


Figure 2-4f

This function is an example of a step function. The name is clear when we examine the graph. A common example of such a function is the cost of mailing a letter. The postage-stamp function, is also an example of a function that requires us to lift the pencil from the paper when we draw its graph. Such a function is usually defined by a set of directions. If we use the formula by which postage on first-class mail is computed, we obtain the graph in Figure 2-4g.

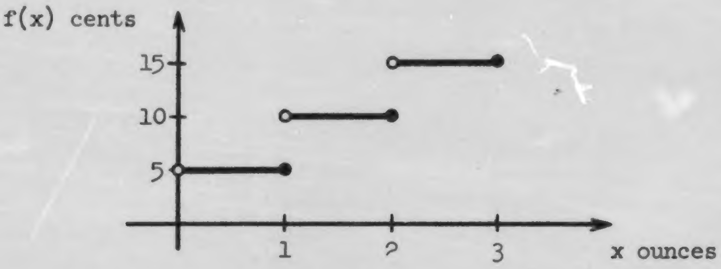


Figure 2-4g

In this instance the domain is the set

$$\{x \mid 0 < x < 320\}$$ and the range

consists of all positive multiples of 5. Less than 320.5

2-5. Second-Degree Relations

Not every equation defines a function. For example when we graph the equation

$$x^2 + y^2 = 25$$

we obtain a circle containing the two points (3,4) and (3,-4). To be a function each first element must be associated with a unique second element. We can obtain two equations expressing y in terms of x that define functions

$$y = \sqrt{25 - x^2}$$

$$y = -\sqrt{25 - x^2}$$

and give a corresponding graph. See Figure 2-5a.

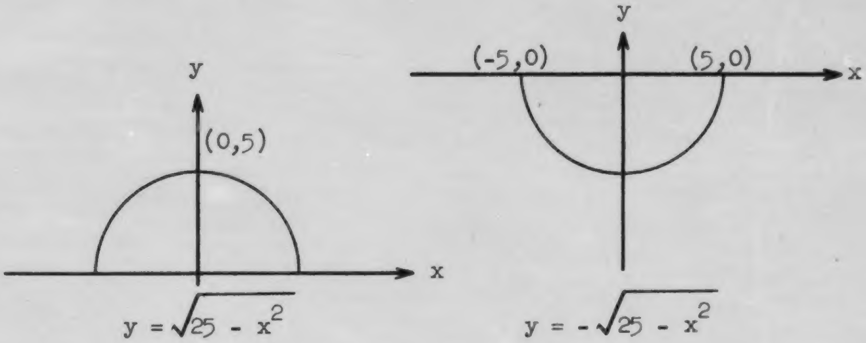


Figure 2-5a

An equation of the second degree in x and y of the form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

($A,B\ldots F \in R$; A,B and C not all three zero) defines a relation, that is, a set of ordered pairs (x,y). Such an equation gives a graph, if one exists, that may be an ellipse, a parabola or a hyperbola according to the following rules:

1. If $B^2 - 4AC < 0$, the curve is an ellipse.
2. If $B^2 - 4AC = 0$, the curve is a parabola.
3. If $B^2 - 4AC > 0$, the curve is a hyperbola.

Remark: The rules as stated are not strictly true. There are a few exceptional cases that produce an isolated point, two intersecting lines, two parallel lines, or a single line. In order to fit even these to the above rules, they are sometimes referred to as degenerate cases, the point being a degenerate ellipse, the two intersecting lines a degenerate hyperbola, and the other two being degenerate parabolas.

Example. Describe the curve determined by the equation

$$x^2 + y^2 - 6x + 10y + 18 = 0 .$$

Draw the graph.

Solution.

$$B^2 - 4AC = 0 - 4(1)(1) = -4$$

The graph is an ellipse or a circle. Since $A = C$ we know it is the latter.

$$\begin{aligned}
x^2 + y^2 - 6x + 10y + 18 &= 0 \\
(x^2 - 6x \qquad) + (y^2 + 10y \qquad) &= -18 \\
(x^2 - 6x + 9) + (y^2 + 10y + 25) &= -18 + 9 + 25 \\
(x - 3)^2 + (y + 5)^2 &= 16
\end{aligned}$$

The center of the circle is the point $(3,-5)$; the radius is 4. See Figure 2-5b.

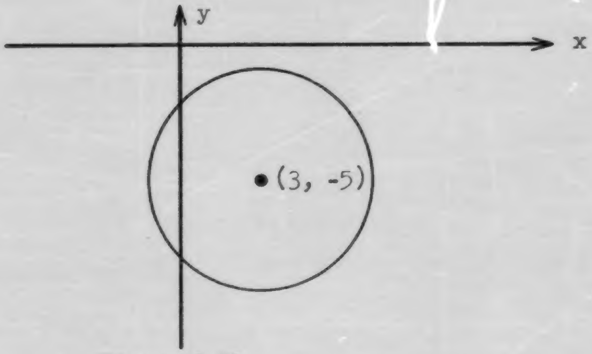


Figure 2-5b

Example. Describe the curve determined by the equation

$$3x^2 - 8x + 4y + 2 = 0 .$$

Draw the graph.

Solution.

$$B^2 - 4AC = 0 - 4(3)(0) = 0 .$$

The graph is a parabola .

$$(3x^2 - 8x \qquad) = -4y - 2$$

$$(x^2 - \frac{8}{3}x + \frac{16}{9} \ ) = - \frac{4}{3}y - \frac{2}{3} + \frac{16}{9}$$

$$(x - \frac{4}{3})^2 = \frac{4}{3}(y - \frac{5}{6})$$

The vertex of the parabola is at the point $(\frac{4}{3},\frac{5}{6})$; the focus is the point $(\frac{4}{3},\frac{1}{2})$; the directrix, the line $y = 1\frac{1}{6}$. See Figure. 2-5c.

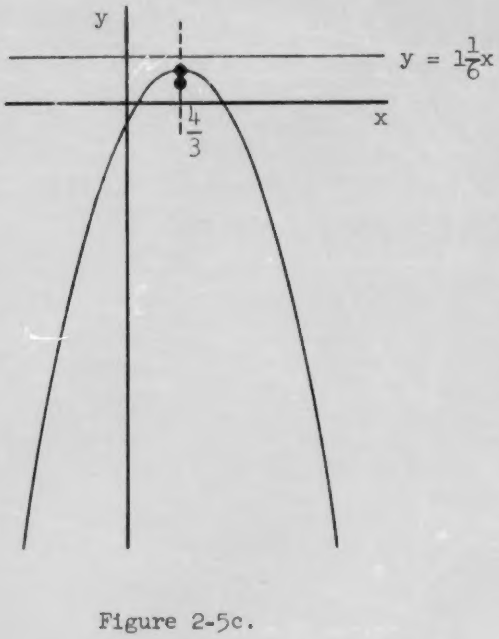


Figure 2-5c.

Chapter 3

THE DERIVATIVE

3-1. A Problem

It is in the nature of the human enterprise to try to get the best of everything: a manufacturer seeks the smallest unit cost for his product, a student tries to complete his homework assignment in the shortest possible time, a demagogue expounds the political philosophy which he believes will garner the greatest number of votes. It is seldom clear what must be done to get the best value. Here we shall develop a systematic attack on a class of these problems.

While the class of "best value" problems treated by the methods of elementary calculus is quite broad, we are only making a beginning in an area which is still a lively field of investigation.

Consider the following problem which a writer faced recently in moving his household goods. The cost of shipping books by parcel post happened to be much lower than the cost of shipment by interstate van. The post office places restrictions on the size of packages: the length plus the girth must not exceed 72 inches. Since there were a great many books, to keep the effort of packing to a minimum the writer sought the largest possible boxes complying with the post office requirement. Assuming the ends of the box to be square[1], what are the dimensions of the box of largest size?

To solve this problem we must know that the post office defines the girth of the box as the perimeter of an end face. We let x denote the number of inches on the side of the square end and y the number of inches on the long dimension of the box; we require that

$$y + 4x = 72 .$$

Under this condition we attempt to maximize the number of cubic inches of volume, V, of the box, where

---

[1] It is not hard to prove that the best box has square ends, but we shall postpone the argument for the sake of brevity here.

49

$$V = x^2y.$$

Setting $y = 72 - 4x$ in the expression for V we obtain

$$V = x^2(72 - 4x).$$

Getting away from the specific details, we see that what we have accomplished is to reduce the problem to the study of the properties of a function $f: x \longrightarrow V$. Our problem is not so much to determine the largest value V_{max} in the range of the function, although that information may also be useful, but to find a value of a in the domain for which $f(a) = V_{max}$. (The domain here consists of those values of x for which the problem is meaningful; tha is, the values between 0 and 72/4.) In order to get some feeling for the problem we may sketch the graph of f by plotting a few easily calculated points and drawing a smooth curve through them. See Figure 3-1a.

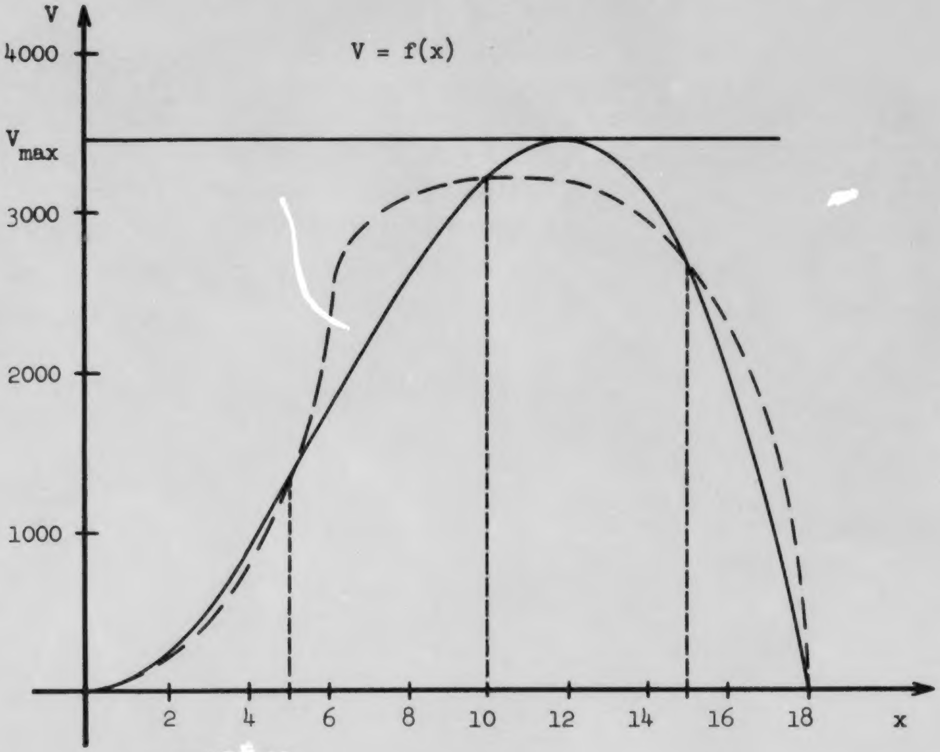


Figure 3-1a. Graph of the function $x \longrightarrow x^2(72 - 4x)$

In this way, we might locate a peak of the graph approximately and we do get some precise information such as $V_{max} \geq f(10) = 3200$. No matter how much information we get this way we shall always be somewhat dissatisfied. In the first place, we have exact information about the function at only a number of calculated points so that even if we happened upon the maximum we might not be aware of it. In the second place, the idea of drawing a smooth curve through the calculated points has limitations. For example, in Figure 3-1a without further calculation we could not be sure that the continuously drawn curve more reasonably represents the function than the broken curve. Furthermore, we cannot eliminate this kind of ambiguity completely by calculating more points. One of our objectives is to devise systematic methods for resolving these difficulties.

Thinking of the problem in visual geometrical terms, we see that the condition for a maximum, $f(a) = V_{max}$, means that the graph of f cannot cross over the horizontal line through $\big(a, f(a)\big)$. The direction of the graph at $\big(a, f(a)\big)$ must therefore also be horizontal, for if the graph met the line at an angle, the two would have to cross. Intuitively, then, the meeting of the line and the graph of the function is a grazing contact; the line is tangent to the graph. To locate a peak of the graph we seek it among the points where the graph has a horizontal tangent. To make some general use of this geometrical idea we express it numerically so that it may serve as a basis for computation. Observing that the direction of the tangent can be represented numerically by its slope, we reformulate our idea: at a peak of the graph the slope of the tangent is zero. We introduce a new function $x \longrightarrow f'(x)$ where $f'(x)$ is the slope of the graph of f at the point $\big(x, f(x)\big)$. If there is a peak of the graph of f at $\big(a, f(a)\big)$ then $f'(a) = 0$. To locate a peak, then, we look among the zeros of $f'(x)$. The function f' is called the derivative of f and the slope of the tangent $f'(x)$, at $\big(x, f(x)\big)$ is called the derivative of f at x.

By now it may seem that we are very far from our second problem and even farther from our initial problem. Let us summarize that which has been accomplished. We have replaced the second problem, about which we knew very little, with a problem about which we know a great deal: to locate a peak of one function we look among the zeros of another function (called the derivative function). The line of approach may seem devious and unproductive. We shall see that it will be fruitful. The discovery of such an avenue of investigation is not beyond the powers of ordinary mortals. Whenever we become unduly impressed by the ingenuity and power of mathematical methods, we should reflect that an investigator will try not one but many approaches. To his admiring audience he will present the one idea that worked and never mention the failures that filled his waste basket with reams of paper. In fact, in this problem we have already briefly considered and rejected one idea, that of finding the maximum value of $f(x)$ by examining a number of its values.

We shall solve our best value problem in Chapter 6 after we develop necessary equipment. It should be stated that the method of solution we rejected was a perfectly practical one. We might have proceeded by calculating values and come very close to the optimum solution.[1] Since, however, problems of this kind arise often it pays to devote some attention to refined methods of solution. Similarly, if we wished to make just one pin we would be content to do it by hand, but if we wished to produce pins by the million we should put a great deal of effort into designing suitable machinery for the purpose. We shall reach the point of view from which the solution of our present problem will appear no more consequential in the light of the methods we shall develop than the production of a single pin in the operation of a pin factory.

Turning back to our initial problem we find that we have so far only replaced it by new problems. In particular, we have not clearly defined the direction of the graph at a given point and, hence, the slope of the tangent. Furthermore, even if the slope of the tangent or derivative is defined at a point there remains the problem of describing the function f' in terms suitable for calculating the solution of the problem.

The art of the mathematician is to extract patterns. A theorem that has a variety of applications is more powerful than one that is useful only in one particular instance. In Chapter 1 a single pattern occurred in the problem involving the direction of a curve and the one concerning instantaneous velocity. In the first we consider the ratio

$$\frac{y_1 - y_0}{x_1 - x_0}, \quad x_0 \neq x_1 ;$$

---

[1] Since the graph is nearly horizontal in the neighborhood of a peak the penalty for missing the exact location of the peak can be expected to be quite small. We shall return to this point later in the text.

in the second, we worked with the ratio

$$\frac{P_1 - P_0}{t_1 - t_0}, \quad t_0 \neq t_1 .$$

Since y is a function of x in the case of the curve and since the position p of the jet-liner is a function of time, we have one pattern for both quotients. The numerator consists of a difference between the value of a function, at two elements of its domain, the denominator, a difference between the same two elements of the domain. Both have the same pattern

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Occurring in a wide variety of applications from jet-liners to post office problems, the difference quotient of this type always defines the average rate of change of a function. For the remainder of this chapter our task will be to refine and extend the difference quotient concept.

Exercises 3-1

1. Explain why in the "post office" problem the domain of the model function $f: x \longrightarrow x^2(72 - 4x)$ is $0 \leq x \leq 18$.
2. Sketch a graph of the function $f: x \longrightarrow x^2(72 - 4x)$. Does your sketch suggest that the authors' observation that

$$V_{max} \geq f(10) = 3200$$

is true? Which of the curves in Figure 3-1 does your sketch most nearly resemble?

3. Summarize the sequence of problems posed in the tack we have taken to solve the initial "post office" problem.

3-2. Numerical Computation of Slope

One of the simplest functions that does not have a straight line graph is the function $f: x \longrightarrow x^2$. As was tacitly taken for granted in the post office problem, we shall assume that the direction of the curve at any point can be described by a straight line. See Figure 3-2. Such a straight line through a point of the curve we shall call a "linear approximation" to the curve at the point in question.

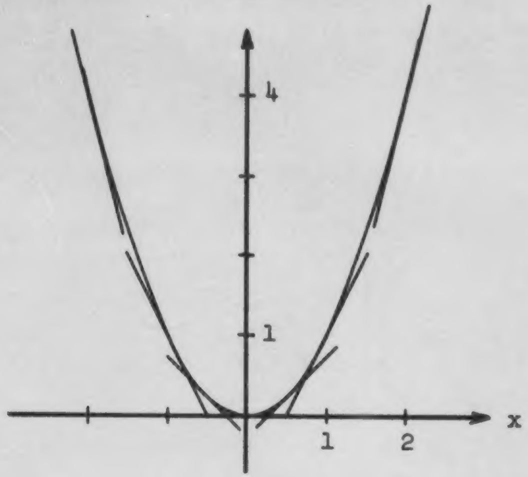


Figure 3-2

Let us start by finding, at the point $(1,1)$, a linear approximation to the curve. We do this by taking a second point close by such as $(1.1, 1.21)$, or $(0.9, 0.81)$. The line through the two points $(1,1)$ and $(1.1, 1.21)$ has slope

$$\frac{1.21 - 1}{1.1 - 1} = \frac{.21}{.1} = 2.1,$$

while the line through the two points $(1,1)$ and $(0.9, 0.81)$ has slope

$$\frac{1 - .81}{1 - .9} = \frac{.19}{.1} = 1.9 .$$

It is reasonable to conclude that any line through the point $(1,1)$ with slope m, such that

$$1.9 < m < 2.1,$$

will be a good linear approximation to the curve. We can squeeze m into a smaller interval by taking points closer to $(1,1)$. If we use $(1.01, 1.001)$ and $(.99, .9801)$, we have one line with slope 2.01 and a second with slope 1.99. The linear approximation has now become a line through point $(1,1)$ and slope m such that

$$1.99 < m < 2.01 .$$

By using points sufficiently close, we can make the interval containing m arbitrarily small. We claim that, for any two points sufficiently close to (1,1), the slope m of the corresponding line will satisfy

$$|m - 2| < \epsilon,$$

where ϵ represents any small positive number arbitrarily chosen. Henceforth we shall call the line, given by the equation

$$(y - 1) = 2(x - 1),$$

with slope 2 through the point (1,1) the best linear approximation to the curve $y = x^2$ at that point. This best linear approximation we shall also hereafter call the tangent to the curve at that point. The statement that the slope of the curve at the point (1,1) is 2 shall be interpreted to mean that the tangent to the curve at that point is 2.

A similar process can be used to find the slope of the curve $y = x^2$ at any other point on the curve. Table 3-2 summarizes the computations approximating to the value of the slope at the point (2,4). It also charts the path for squeezing in on the value of the slope at the point (-1.5, 2.25).

x_0,y_0	x_1,y_1	m_1	m_2	m
(2,4)	(2.1, 4.41)	4.1		3.9 < m < 4.1
(2,4)	(1.9, 3.61)		3.9	
(2,4)	(2.01, 4.0401)	4.01		4.01 < m < 3.99
(2,4)	(1.99, 3.9601)		3.99	
(-1.5, 2.25)	(-1.4, 1.96)	?		? < m <
(-1.5, 2.25)	(-1.6, 2.56)		?	?
(-1.5, 2.25)	(-1.49, ?)	?		? < m <
(-1.5, 2.25)	(?, ?)		?	?

Table 3-2

From the information in Table 3-2 we suspect that for all points sufficiently close to (2,4) the difference between 4 and the slope of the corresponding line is arbitrarily small:

$$|m - 4| < \epsilon .$$

Hence the best linear approximation to the curve $y = x^2$ at the point $(2,4)$ seems to be the line given by the equation

$$(y - 2) = 4(x - 4) .$$

This line we call the tangent and we say the slope of the curve is 4.

Exercises 3-2

1. Complete the computations for Table 3-2 to find a set of approximations which suggest the slope of the curve $y = x^2$ at the point $(-1.5, 2.25)$. What is the equation of the tangent to the curve at this point?

2. By assembling a table similar to Table 3-2, determine the slope of the curve $y = x^2$ at the points

   (a) $(3,9)$
   (b) $(-1.25, 1.5625)$
   (c) $(0,0)$

3. By a similar procedure, determine the slope of the curve $y = x^3$ at the points

   (a) $(1,1)$
   (b) $(2,8)$
   (c) $(-1.1, -1.330)$

4. Using a similar procedure to that which you employed in exercises 1 through 3 can you find the slope of the tangent through the graph of $x \rightarrow x^3$ at the origin?

5. In the text it is asserted, for the curve $y = x^2$ at the point $(1,1)$, that for points sufficiently close to $(1,1)$ the slope m of the corresponding line will satisfy

$$|m - 2| < \epsilon$$

   for all positive values of ϵ. Find two points close to $(1,1)$ so that

   (a) $|m - 2| < .1$
   (b) $|m - 2| < .001$
   (c) $|m - 2| < .0004$

3-3. *The Slope Function for $y = x^n$*

We have seen how we can, through numerical computation, find the slope of a curve. *Such a method is tedious, as you may know!* Let us generalize our procedure and search for a pattern that can be broadly applied.

We shall use the curve $y = x^2$ again, but let us consider a fixed point (a,a^2) and two points close by (x_0,x_0^2) and $(x_1x_1^2)$ where

$$x_0 < a < x_1$$

and

$$|a - x_0| = |a - x_1| .$$

Also we must assume that the segment of the curve that contains these three points is a continuous smooth curve, that is a segment that has no *breaks*, or *sharp turns*.

The slope of the line through (a,a^2) and (x_0,x_0^2) is

$$\frac{a^2 - x_0^2}{a - x_0} .$$

The line through (a,a^2) and (x_1,x_1^2) has slope

$$\frac{a^2 - x_1^2}{a - x_1} .$$

Now m, the slope of the linear approximation or tangent is

$$\frac{a^2 - x_0^2}{a - x_0} < m < \frac{a^2 - x_1^2}{a - x_1} ,$$

or, by factoring,

$$\left(\frac{a - x_0}{a - x_0}\right)\left(a + x_0\right) < m < \left(\frac{a - x_1}{a - x_1}\right)\left(a + x_1\right)$$

Since $|a - x_0| = |a - x_1| \neq 0$, we have

$$a + x_0 < m < a + x_1 .$$

As x_0 and x_1 become closer to a, the sums $a + x_0$ and $a + x_1$ both approximate 2a. By choosing values sufficiently close, we have

$$|m - 2a| < \epsilon .$$

We conclude that the slope of the curve $y = x^2$ at any point (a,a^2) on the curve is 2a. In this case, the slope of the curve is a function of the first coordinate of the point! At (1,1), the slope is 2; at (10,100), the slope is 20; at (-4, 16), -8.

Generally we denote an arbitrary point on the curve $y = x^2$ with the ordered pair $\big(x,f(x)\big)$. Where we use this customary notation, then the slope of the curve at that point will be 2x. We shall call the function that associates the coordinate x with the slope 2x the slope function and denote it by the symbol f'. Thus we have

$$f': x \longrightarrow 2x .$$

One generalization has revealed a great deal; let us explore some more.

Next let us consider the function defined by the equation $y = x^n$ where n is a positive integer. As before, we seek to determine the value of the slope m within an interval that can be made arbitrarily small. If the two boundary values converge on one value, we shall call the latter the slope. We start with a point $\big(a,f(a)\big)$ and two other points close by $\big(x_0,f(x_0)\big)$ and $\big(x_1,f(x_1)\big)$. Again we assume that there are no jumps or breaks in our smooth curve and that

$$x_0 < a < x_1 .$$

Proceeding as before, m will now be defined by the inequality

$$\frac{f(x_0) - f(a)}{x_0 - a} < m < \frac{f(x_1) - f(a)}{x_1 - a}$$

or using the fact that $f(x) = x^n$

$$\frac{x_0^n - a^n}{x_0 - a} < m < \frac{x_1^n - a^n}{x_1 - a} .$$

58

Recall that any binomial of the form $x^n - a^n$ can always be factored into two factors, one of which is $x - a$. For example

$$x^1 - a^1 = (x - a)(1)$$
$$x^2 - a^2 = (x - a)(x + a)$$
$$x^3 - a^3 = (x - a)(x^2 + xa + a^2)$$
$$x^4 - a^4 = (x - a)(x^3 + x^2a + xa^2 + a^3)$$
$$\vdots$$
$$x^n - a^n = (x - a)(x^{n-1} + x^{n-2}a^1 + x^{n-3}a^2 + \dots + a^{n-1}) .$$

Substituting accordingly in each numerator, and then simplifying, we have

$$x_0^{n-1} + x_0^{n-2}a^1 + x_0^{n-3}a^2 + \dots + a^{n-1} < m < x_1^{n-1} + x_1^{n-2}a^1 + x_1^{n-3}a^2 + \dots + a^{n-1} .$$

By choosing values of x_0 and x_1 sufficiently close to a, we can make the differences

$$x_0^{n-1} + x_0^{n-2}a + x_0^{n-3}a^2 + \dots + a^{n-1} - na^{n-1}$$

and

$$x_1^{n-1} + x_1^{n-2}a + x_1^{n-3}a^2 + \dots + a^{n-1} - na^{n-1}$$

arbitrarily small. Likewise we can make

$$|m - na^{n-1}| < \epsilon$$

by choosing values of x_0 and x_1 sufficiently close to a. We conclude that the slope of the curve $y = x^n$ at any point (a,a^n) is na^{n-1} .

We have not only obtained a rule for finding the slope of a curve $y = x^n$ at some point $\big(x,f(x)\big)$ on the curve but also we have derived a new function. If a function f is defined as $f\colon x \longrightarrow x^n$, where n is a positive integer, then there is another function f' such that $f'\colon x \longrightarrow nx^{n-1}$, when n is a positive integer, and $f'(x)$ is the slope of the curve $y = f(x)$ at the point $\big(x,f(x)\big)$.

Example. Find the slope function of the function $f\colon x \longrightarrow x^3$. Then find the slope of the curve $y = x^3$ at the point whose abscissa is 3.

Solution. If $f: x \longrightarrow x^3$ then $f': x \longrightarrow 3x^2$ then $f'(x) = 3x^2$, and $f'(3) = 27$.

3-4. The Velocity Function

Usually one first meets the notions of speed and velocity through familiarity with an automobile speedometer which reports the speed of the vehicle at each instant of motion. The automobile, intended to go forward in a single direction, has a magnetic device that reports speed, which is one component of velocity. Velocity has magnitude and direction and is denoted by a signed number usually. Speed can be considered as the absolute value of velocity.

Since the distance the normal automobile traverses is rather difficult at best to describe with a rather simple equation, let us consider a situation that can be so described. Suppose a boy, lying on his back, shoots an arrow straight up from the ground with an initial velocity of 64 feet per second. Assuming the arrow is only influenced by gravity, we know from actual experiments that the height, h, of the arrow is a function, g, of time $[h = g(t)]$ defined by

$$h = 64t - 16t^2 .$$

To determine the velocity of the arrow at any particular moment, we start with the notion of average velocity during a period of time. This we define as the quotient

$$\frac{\text{change in distance}}{\text{time interval}} = \frac{f(t_1) - f(t_0)}{t_1 - t_0} .$$

If we let $t_1 = 2$ and $t_0 = 1$, and then use this pattern, we have

$$\frac{f(2) - f(1)}{2 - 1} = [64(2) - 16(4)] \cdot [64(1) - 16(1)] = 16 .$$

Hence we can say that the average velocity of the arrow during the second of travel is 16 feet per second. The average velocity during a second can be quite different from the velocity at a particular moment. Even the concept of a moment suggests a period of time, but a very, very short period. For the velocity during a moment, sometimes called instantaneous velocity, we must use a shorter period of time. Table 3-4 contains the results that come from considering small periods of time near $t = 1$.

t_1	t_2	$t_2 - t_1$	$f(t_2) - f(t_1)$	$\frac{f(t_2) - f(t_1)}{t_2 - t_1}$
1	1.1	.1	3.04	30.4
1	.9	-.1	-3.36	33.6
1	1.01	.01	.3184	31.84
1	.99	-.01	- .3216	32.16

Table 3-4

From Table 3-4, we see that the average velocity during the 0.1 second before $t = 1$ is 30.4 feet per second and that it is 33.6 feet per second during the 0.1 second after $t = 1$. It is reasonable to conclude that the velocity v when $t = 1$ is

$$30.4 < v < 33.6 .$$

If we make the period of time .01 second, we have

$$31.84 < v < 32.16 .$$

By using even shorter periods of time, we can make the interval containing v artibrarily small. No matter what small positive number ϵ is chosen, there is a period of time so brief that

$$|v - 32| < \epsilon .$$

We call 32 feet per second the velocity at $t = 1$.

As with the slope function, we can generalize our procedure and obtain a velocity function for the arrow. Let us consider a fixed time t_1 and a time t very close to t_0 (t may be either greater or less than t_1). Our difference quotient expressing average velocity is

$$\frac{(64t - 16t^2) - (64t_0 - 16t_0^2)}{t - t_0} = \frac{64(t - t_0) + 16(t_0^2 - t^2)}{t - t_0} = 64 + 16(t_0 + t) .$$

[Note: If we considered t' and t'', $t' < t_0 < t''$, we would now have

$$64 + 16(t_0 + t') < v < 64 + 16(t_0 + t'').]$$

By choosing a time t sufficiently close to t_0, we have, no matter how small a positive number ϵ may be,

$$|(t_0 + t) - 2t_0| < \epsilon .$$

Consequently we can say that the velocity, v, at any time, t_0, in the domain of the function g is defined by the equation

$$v = 64 - 16(2t_0) = 64 - 32t_0 .$$

Hence v is a function of t also. Since $h = g(t)$, we write $v = g'(t)$.

Exercises 3-4

1. If the distance s of an object is given by the equation $s = 16t^2 + 25$, find the average velocity during the period of time from $t = 2$ to $t = 3$.
2. Using the situation in exercise 1, find through a numerical computation the velocity when $t = 2$.

   (Additional exercises of this type are found in a supplementary section submitted in the Commentary for Teachers.)

3-5. The Derivative

By the name, the slope function is attached to a curve and the velocity function is attached to motion. Suppose we broaden our interpretation by considering the continuous function f defined by an equation $y = f(x)$ Let a be any element in the domain of the function and x be a second element $(a \neq x)$. We call

$$\frac{f(x) - f(a)}{x - a}$$

the difference quotient and it is the average rate of change of the function from a to x. Now we proceed to make the difference $x - a$ smaller and smaller. If the difference quotient thereby becomes arbitrarily close to some number L, that is

$$\left|\frac{f(x) - f(a)}{x - a} - L\right| < \epsilon ,$$

then the number L is called a limit, and is the derivative of f at a.

Intuitively we speak of the difference quotient approaching some definite value as a limit as x approaches a, symbolized by $x \longrightarrow a$.

DEFINITION 3-5. The derivative of f at a, written $f'(a)$, is given by

$$f'(a) = \lim_{x \longrightarrow a} \frac{f(x) - f(a)}{x - a} .$$

This may also be written as $\lim_{a} \frac{f(x) - f(a)}{x - a}$.

There are several remarks to be made about this definition. First and foremost, the definition is without meaning when $x = a$. (Why?) Second, the definition does not specify or indicate what the number $f'(a)$ actually is. Third, assuming that the value of the limit is unique, we have each number a in the domain of the function f associated with a number $f'(a)$. Thus we have a function $f': x \longrightarrow f'(x)$ which we call the derived function. Fourth, inherent in the definition is the idea of two limits: one the result of considering $x < a$ (occasionally referred to as the left-hand limit); the other, the result of considering $x > a$ (the right-hand limit).

The subject of differential calculus is devoted to the consequences and applications of this definition. Two applications, the slope function and the velocity function, we have met and we can now make a definition accordingly.

DEFINITION 3-5a. The tangent to the graph of the function f, defined by the equation $y = f(x)$, at the point $\big(a, f(a)\big)$ is the line

$$y - f(a) = f'(a)[x - a] .$$

DEFINITION 3-5b. The velocity v at time t_0 of an object whose displacement s is a function of t, $\big(s = f(t)\big)$, is $v_{t_0} = f'(t_0)$.

The notation that we have used for the derivative of f at a, $f'(a)$, is an adaptation of Newton's. Another notation is $D_x y$, read "the derivative of y with respect to x," which is useful when we have a function defined by an equation such as $y = x^2$. We write $D_x y = 2x$, meaning the derived function defined by the equation $y = 2x$.

There are also alternate forms of the definition. To form the difference quotient, we used two elements in the domain of the function a and x $(a \neq x)$. If we choose to designate two elements in the domain as x and $x + h$, where h is a non-zero number, then the difference quotient is

$$\frac{f(x + h) - f(x)}{h}$$

and we have

$$f'(x) = \lim_{h \longrightarrow 0} \frac{f(x + h) - f(x)}{h} .$$

On occasion, it is useful to designate the two elements in the domain as x and $x + \Delta x$, where Δx is a non-zero number called the increment of x. Assuming that we have a function defined by an equation $y = f(x)$, we let $f(x + \Delta x) - f(x) = \Delta y$. Then the difference quotient is $\frac{\Delta y}{\Delta x}$ and the derivative is

$$\lim_{\Delta x \longrightarrow 0} \frac{\Delta y}{\Delta x} .$$

Exercises 3-5

1. If we have a function g defined by an equation $s = g(t)$, give three alternate forms of the derivative using this notation.

Chapter 4

LIMITS

4-1. Introduction

One of the most fundamental ideas underlying every aspect of calculus is the limit concept. It is easy to accept without questioning the statement "The value of $x + 2$ approaches 5 as x approaches 3." The meaning is not obscure; the statement is paraphrased by students in about the same way. When we modify the statement, however, and assert "The limit of $x + 2$ is 5 as x approaches 3", the meaning is no longer uniformly clear. Students interpret the latter statement in an assortment of ways, some of which are incorrect. The difficulty arises from the word limit. For example, if the statement is converted to the question "What is a reasonable value to assign to the expression $x + 2$, as x approaches 3?" less difficulty ensues. Our problem at the moment is to define limit in such a reasonable way that it will provide a uniform interpretation to the simple statement given above and also give ultimately the ability to answer such questions as

(a) What is the limit of the expression $\frac{\sqrt{2 + h} - \sqrt{2}}{h}$ as h approaches zero?

(b) What is the limit of the expression $\frac{\sin x}{x}$ as x approaches zero?

(c) What is the limit of the expression $\frac{x^2}{2x + x^2}$ as x approaches infinity?

Intuition, which we have relied upon happily up to now, does not give ready or uniform answers to these questions. Nor does intuition provide the tools to prove rigorously some of the earlier assertions, particularly if we divorce geometric interpretations from our discussion. In Chapters 1 and 3 the word limit was interpreted by differences. This will prove to be a fruitful course as we now learn.

4-2. Limit of a Sequence

Since the notion of a limit of a sequence is one we met when we studied geometric series, we start with this notion. It serves as a good introduction to the idea of limit of a function, which is the rock on which calculus is built.

Sequences are functions, but since they are functions that associate with each positive integer, n, a number a_n where $a_1 = f(1)$, $a_2 = f(2)$, etc., they are called discrete functions rather than continuous functions. The numbers a_1, a_2, a_3, ... are called the terms of the sequence, $\{a_n\}$. Sequences may be of various types. There are finite sequences, for example

(a) $-1, -\frac{1}{2}, 0, \frac{1}{2}$. ($a_n = \frac{n-3}{2}$ for $n = 1, 2, 3, 4$)

(b) $6, -12, 18, -24, 30, -36$. ($a_n = (-1)^{n+1}(6n)$ for $n = 1, \ldots, 6$).

There are infinite sequences such as

(c) $1, 3, 5, 7, \ldots$ ($a_n = 2n - 1$).

(d) $\frac{3}{2}, \frac{5}{4}, \frac{9}{8}, \frac{17}{16}, \ldots$ ($a_n = \frac{2^n + 1}{2^n}$) .

Also there are sequences consisting of successive partial sums of a series. (Frequently such sequences are denoted $\{S_n\}$ although this distinction is not necessary when the content of the problem is sufficient to make it clear.) For example, from the series $\frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots$, we have a sequence of partial sums

(e) $1, \frac{3}{2}, \frac{7}{4}, \frac{15}{8}, \frac{31}{16}, \ldots$ ($S_n = \sum_{k=1}^{n} \frac{1}{2^{k-1}}$)

where $S_1 = 1$

$S_2 = 1 + \frac{1}{2}$

$S_3 = 1 + \frac{1}{2} + \frac{1}{4}$

Frequently we write

$$(f)\quad S_n = \sum_{k=1}^{n} \frac{1}{k}$$

to mean the sequence of partial sums $S_1 = 1$, $S_2 = \frac{3}{2}$, $S_3 = \frac{11}{16}$, ... that evolves from the series $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} \dots$.

The most interesting characteristic of sequences for us now is that some sequences, said to be convergent, approach some finite value L, called the limit, as the number of terms increases. Some sequences do not approach a finite value L as n becomes progressively larger; such sequences do not have a limit and are called divergent.

It is difficult, by intuition alone, to distinguish between convergent and divergent sequences. For example, it may be hard to classify examples e and f above. Moreover, it is important to have a precise definition that will lend itself to a proof. Let us consider examples c and d given earlier. The sequence 1, 3, 5, 7, ... clearly does not approach one finite value. Whatever number might be named, for example, 10,437,321, members of the sequence can be demonstrated that are very much larger. (Sometimes it is said that the limit is infinity which implies that there is no finite limit, the only kind we are considering. The concept of limit has enough ambiguity without compounding this vagueness with infinity!) The sequence (d) $\frac{3}{2}, \frac{5}{4}, \frac{9}{8}, \frac{17}{16}, \dots$ on the other hand, seems to be approaching closer and closer to 1 as n gets larger and larger. We guess that this sequence has a limit, namely 1, and we write $\lim_{\infty} a_n = 1$. To interpret this statement, we turn naturally to differences which are suggested by the use of the comparatives, closer and larger. When we assert that the limit is 1, we mean that the difference $a_n - 1$ can be made as small as we desire by choosing n sufficiently large. (The symbol ∞ means only that regardless of what number n we might specify, however large, there are larger numbers.) We still have a rather vague and subjective statement. What is arbitrarily small to one person or at one time may not be so small to another person or upon another occasion. Do we mean that

$$|a_n - 1| < .01\text{[*]}, \quad \text{or} \quad |a_n - 1| < .00001\text{[*]}, \quad \text{or} \quad |a_n - 1| < .0000001\text{[*]}?$$

If indeed the latter is small enough to satisfy even the most small-minded person, then when $n = 24$ the difference will be that small. Since $a_n = \frac{2^n + 1}{2^n}$, $|a_n - 1| = \frac{1}{2^n}$ and $\frac{1}{2^{24}} < .0000001$. It is important to note that for all $n > 24$, such as $n = 32$ or $n = 108$, the condition $|a_n - 1| < .0000001$ is also satisfied. But we shouldn't be involved in a discussion of small-minded people. Let's broaden the discussion. We recognize that smallness is relative and that individual instances of small numbers do not suffice. To have a proof we must argue from a general case. Let us consider $|a_n - 1| < \epsilon$ where we allow ϵ to be any arbitrarily small positive number. Since we know that $|a_n - 1| = \frac{1}{2^n}$, we recognize that we can generate a difference less than any ϵ by raising 2 to a suitable power, that is, by using a value of n suitably large. The particular number n that is suitably large we designate by N. We write $n \geq N$, where the selection of N is dependent upon the choice of ϵ. If $|a_n - 1|$ is to be less than , it also means $\frac{1}{2^n}$ is to be less than ϵ. To determine a value of n that makes this statement true, we solve the inequality $\frac{1}{2^n} < \epsilon$ for n. Thus

$$\frac{1}{2n} < \quad \Longleftrightarrow \frac{1}{\ } < 2^n \Longleftrightarrow \log \frac{1}{\ } < \log 2^n \Longleftrightarrow \frac{-\log \epsilon}{\log 2} < n .$$

Whenever n is greater than $\frac{-\log \epsilon}{\log 2}$, which is a positive number whenever $\epsilon < 1$ (Recall $\log 1 = 0$), the inequality $|a_n - 1| < \epsilon$ is true. Thus if we let $N = .\left|\frac{\log \epsilon}{\log 2}\right|$, we can assert that $|a_n - 1| < \epsilon$ for all $n > N$.

In our proof, we are not compelled to demonstrate the smallest N. Consequently, since $\log 2 > .2$, we know that

---

[*] A particle .01 inches in diameter irritates your eye.

[*] A particle .00001 inches in diameter can not be seen by the human eye.

[*] A particle .0000001 inches in diameter is the smallest mass discernible by an electron microscope.

$$\frac{-\log \epsilon}{\log 2} < \frac{-\log \epsilon}{.2} = -5 \log \epsilon \ .$$

Hence we can also assert that $|a_n - 1| < \epsilon$ for all $n \geq N$ where $N = |5 \log \epsilon|$.

The foregoing arguments constitute proofs because we have a general method for determining N for all possible choices of ϵ.

Our statements about ϵ and N have a geometric interpretation. Let us plot the members of $\{a_n\}$ where $a_n = \frac{2^n + 1}{2^n}$ on a number line. See Figure 4-2a. To avoid confusion we label the points associated with the elements of the sequence $a_n = \frac{2^n + 1}{2^n}$ by the symbols that represent them in the sequence.

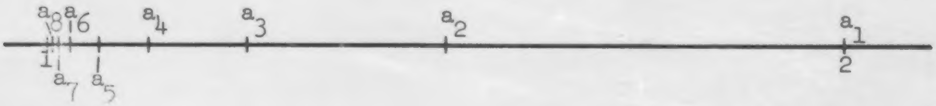
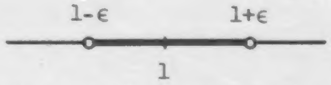


Figure 4-2a.

We see that for larger values of n, the points cluster about 1. When we choose ϵ we establish an interval



(In our example we actually are concerned with the right-hand half of the interval only even though the use of the absolute value notation $|a_n - 1| < \epsilon$ allows both.) and then through our choice of N, we indicate the means of placing all terms a_n for $n \geq N$ within the interval. The choice of specific ϵ may vary and, accordingly, the choice of N must vary. By solving for N in terms ϵ, however, we have the appropriate value for N whatever ϵ may be.

Let us look at another sequence, $\{a_n\} = 4 + \frac{(-1)^{n+1}}{n}$. We have $5, 3\frac{1}{2}, 4\frac{1}{3}, 3\frac{3}{4}, 4\frac{1}{5}, \ldots$. In this case, we can guess that the limit of the sequence is 4. To prove this, we must demonstrate N such that $|a_n - 4| < \epsilon$ whenever $n \geq N$. Since $|a_n - 4| = \frac{1}{n}$, we have $\frac{1}{n} < \epsilon$ which is equivalent to $\frac{1}{\epsilon} < n$. Now if we let $N = \frac{1}{\epsilon}$, we have completed our proof that the limit is 4. Let us give a geometric interpretation to these ideas.

If we plot the numbers $\{a_n\}$, we have Figure 4-2b.

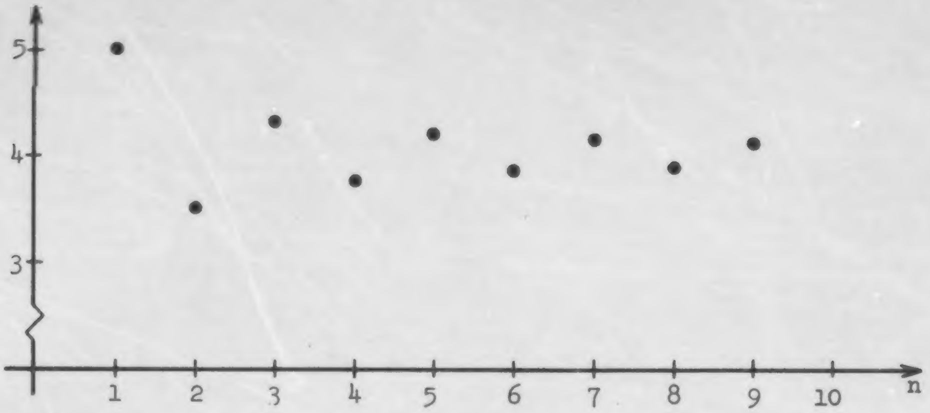


Figure 4-2b.

When we use the coordinate plane, the statement $|a_n - 4| < \epsilon$ establishes an interval on the vertical axis and a corridor on the plane

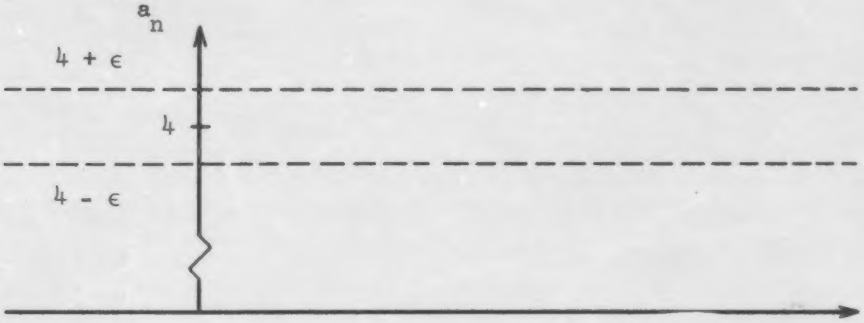


Figure 4-2c.

which is 2ϵ in width. See Figure 4-2c. Now we must indicate a number N such that whenever $n \geq N$, a_n will be within the corridor. And we must be able to produce this N for any value of ϵ whatsoever. The choice of N establishes the left hand boundary of the corridor. See Figure 4-2d. The values of ϵ and N establish a closed corridor which contains all a_n for $n \geq N$. Note that any value N_1, such that $N_1 > 1$, also establishes a closed corridor that contains all members of the sequence a_n for $n \geq N_1$.

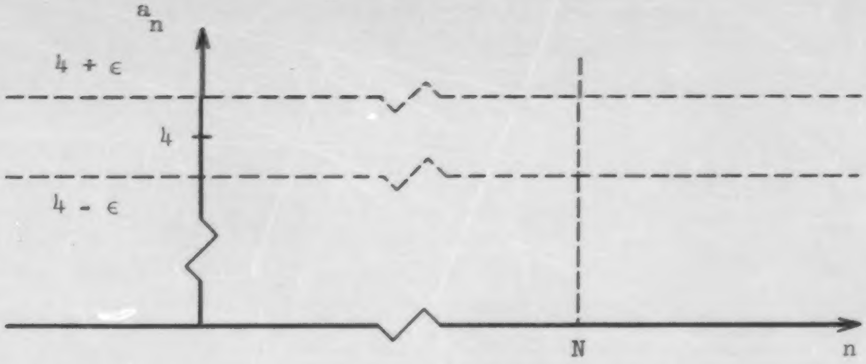


Figure 4-2d.

As mentioned earlier, some sequences do not converge. For example, the sequence 1, -1, 1, -1, 1, ..., where $a_n = (-1)^{n+1}$ does not. There is no number L such that the difference $|a_n - L|$ is always less than some small positive number.

It is not quite so easy to see that the sequence $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$ is divergent. The partial sums are

$$\begin{aligned}
S_1 &= 1 && = 2(\tfrac{1}{2}) \\
S_2 &= 1 + \tfrac{1}{2} && = 3(\tfrac{1}{2}) \\
S_4 &= (1 + \tfrac{1}{2}) + (\tfrac{1}{3} + \tfrac{1}{4}) && > 4(\tfrac{1}{2}) \\
S_8 &= (1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4}) + (\tfrac{1}{5} + \ldots + \tfrac{1}{8}) && > 5(\tfrac{1}{2}) \\
S_{16} &= (1 + \ldots + \tfrac{1}{8}) + (\tfrac{1}{9} + \ldots + \tfrac{1}{16}) && > 6(\tfrac{1}{2})
\end{aligned}$$

Starting at each member $\frac{1}{n}$, where n is a power of 2, that is $n = 2^m$, each member in the next block of 2^m terms,

$$\frac{1}{2^m + 1} + \frac{1}{2^m + 2} + \ldots + \frac{1}{2^m + 2^m},$$

will be greater than or equal to $\frac{1}{2^m + 2^m}$. Hence this block of terms is greater than or equal to

$$\frac{2^m}{2^m + 2^m} = \frac{1}{2}$$

Since there are an infinite number of such blocks, the sequence has no limit and is therefore divergent.

We are ready to make an important definition.

DEFINITION 4-1. The sequence $\{S_n\}$ is said to have the number L as a limit when n tends to infinity provided that for each positive number ϵ there corresponds an integer N such that

$$|S_n - L| < \epsilon \quad \text{whenever} \quad n > N.$$

We write

$$\lim_{\infty} S_n = L.$$

In our discussion of limits in this section, we have referred to "the limit" as though it were unique. Let us prove this now.

THEOREM 4-2. If a sequence $\{a_n\}$ has a limit, it is unique.

Proof. Suppose a sequence has twc distinct limits L_1 and L_2. Since $L_1 \neq L_2$, $L_1 - L_2$ is an interval. By choosing appropriate values for ϵ_1 and ϵ_2 we can create intervals centered on L_1 and L_2 that do not overlap. See Figure 4-2e.

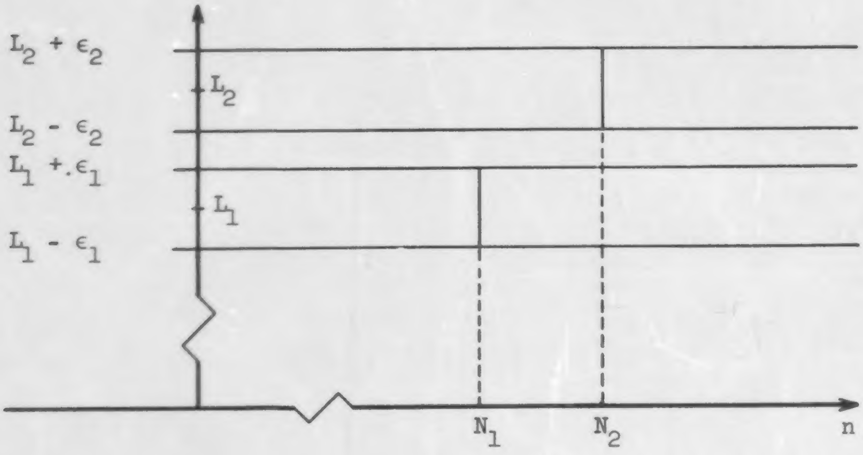


Figure 4-2e.

Since L_1 is a limit, there is some number N_1 such that all members a_n, $n > N_1$, will be in the closed corridor created by the interval and N_1. Also there will be another number N_2 such that all members a_n, $n > N_2$, will be in this second closed corridor. What about the members a_n when $n > N_1$ and $n > N_2$? It is impossible for them to be in two disjoint corridors. Thus we have a contradiction of the hypothesis. Two distinct limits are impossible.

Exercises 4-2

1. Decide which of the following sequences converges. State the limit if the sequence converges and prove the statement.

   (a) $0, 1, 0, 2, 0, 3, \ldots$

   (b) $1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8}, \ldots$

   (c) $-1, 2, -3, 4, -5, \ldots$

   (d) $0.6, 0.66, 0.666, 0.6666, \ldots$

   (e) $2, \frac{5}{2}, \frac{8}{3}, \frac{11}{4}, \ldots$

2. Decide what the limit is for each sequence given below. For the given ϵ, find an integer N.

   (a) $S_n = 2 + \frac{1}{n}$, $\epsilon = .01$

   (b) $S_n = \frac{2n + 3}{n}$, $\epsilon = 0.5$

3. Find the sum of the series

$$\sum_{k=1}^{\infty}\left(\frac{2}{k^2 + 2k + 1} - \frac{2}{k^2}\right)$$

4-3. Limit of a Function of a Continuous Variable

Generally in the study of introductory calculus, we are not concerned with discrete functions such as sequences although they have tremendous significance in more advanced courses. We have discussed the limits of sequences, however, since they serve to introduce rather nicely the concept of a limit of a function whose domain consists of all real numbers throughout an interval. Although we have from the previous chapter one of the most significant limits in the entire realm of mathematics

$$\lim_{a} \frac{f(x) - f(a)}{x - a} ,$$

it is better to start with a simpler illustration, one that we quickly grasp intuitively.

Suppose we have a function, f, defined by the equation $y = 3x - 2$. It is reasonable to assert that y approaches 4 as x approaches 2. In other words, that y can be made to approximate 4 within any margin of error by taking x sufficiently close to 2. We can paraphrase these statements by asserting that the limit of $f(x)$, as x tends to 2, is 4, and succinctly writing

$$\lim_{x \longrightarrow 2} f(x) = 4 .$$

We must note and be careful to remember that this assertion is not the same as saying $f(2) = 4$. Later on we shall discuss the implications arising from the assertion $\lim_{x \longrightarrow 2} f(x) = 4$. For the moment note that the expression mentioned above,

$$\frac{f(x) - f(a)}{x - a} ,$$

does not have meaning when $x = a$.

So far our statements do not provide the means of proving the assertions; to do this we must interpret the words "approximate" and "close to" in terms of differences, or intervals on the number line. If we say that the values of $f(x)$ approach 4, are close to 4, or the limit $f(x) = 4$, we mean that the difference $|f(x) - 4|$ can be made less than ϵ for each arbitrary small positive number ϵ. We assert that this difference can be made less than ϵ by using values of x sufficiently close to 2. The phrase "sufficiently close to 2" is also interpreted as a difference. We assert there is an interval, $2 - \delta < x < 2 + \delta$, or equivalently $|x - 2| < \delta$, such that any value of x within the interval, except for $x = 2$, will give a value for $f(x)$ that lies within the interval $4 - \epsilon < f(x) < 4 + \epsilon$. Note that we now have two distances ϵ and δ to consider. The number ϵ is a measure of the range interval; it represents the margin of error. The number δ is a measure of the domain interval; it represents the control of the error.

Earlier we mentioned the inequality $4 - \epsilon < f(x) < 4 + \epsilon$ can be interpreted as an interval on the number line or as a corridor on the plane. Again this is true. See Figure 4-3a.

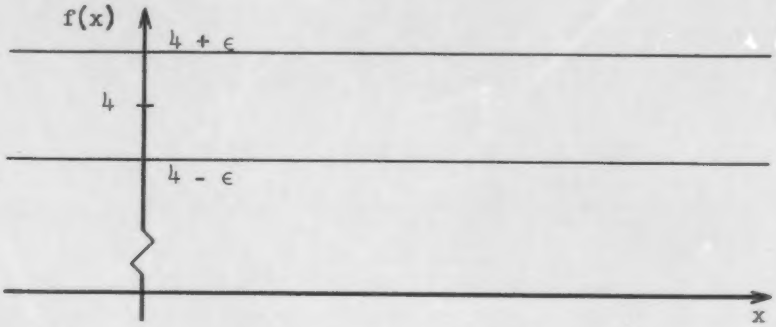


Figure 4-3a.

Now the role of the interval, $2 - \delta < x < 2 + \delta$, is to close off part of the corridor. The interval, $|x - 2| < \delta$, must be so chosen that it walls off all of the domain for which $f(x)$ lies outside the corridor.

Think of $|x - 2| < \delta$ as an interval on the x-axis producing a vertical corridor on the plane. See Figure 4-3b. This second corridor, where it

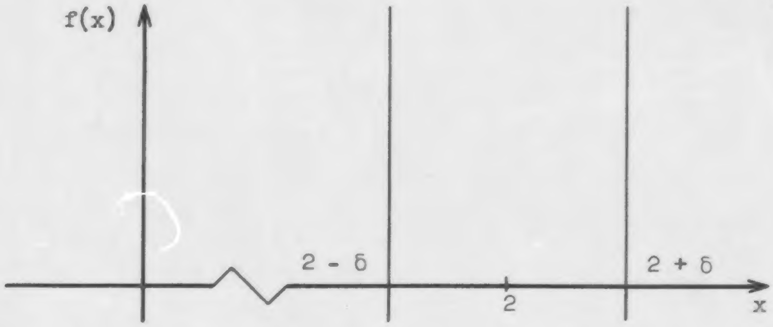


Figure 4-3b

overlaps and closes off a part of the first corridor, $|f(x) - 4| < \epsilon$. See

Figure 4-3c. The graph of f for points when $2 - \delta < x < 2 + \delta$ must lie in the shaded rectangle. If we draw the graph of $f: x \longrightarrow 3x - 2$ and establish an arbitrary corridor 2ϵ in width by choosing an interval, $4 - \epsilon < f(x) < 4 + \epsilon$,

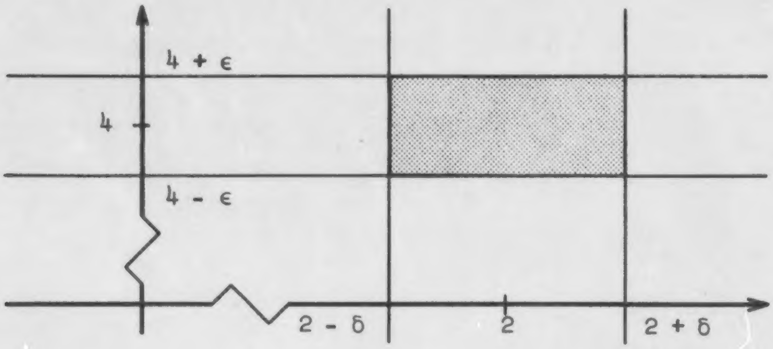


Figure 4-3c.

on the vertical axis, we observe that the corridor intersects the graph at two points P_1 and P_2. See Figure 4-3d. The vertical corridor must be so chosen by the proper designation of δ that this second corridor intersects the graph between points P_1 and P_2. See Figure 4-3e. Note well that the

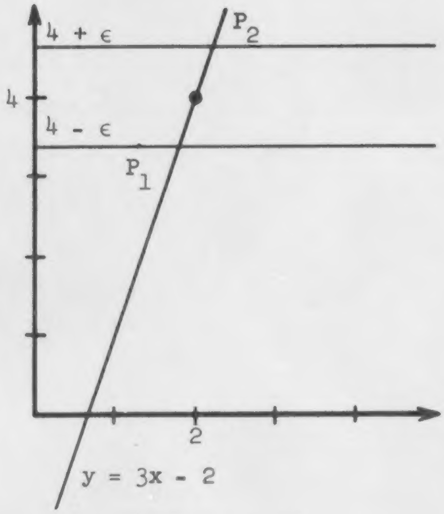


Figure 4-3d.

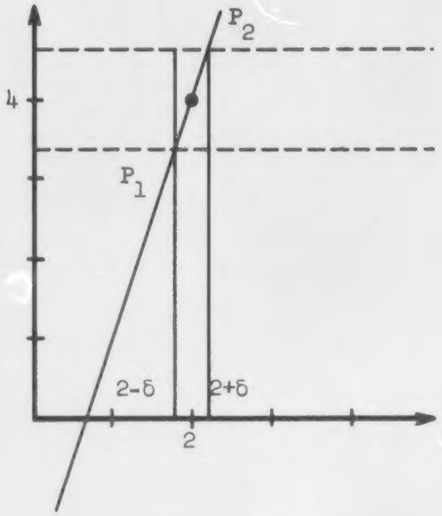


Figure 4-3e.

value of δ is dependent upon the value chosen for ϵ. The essence of any discussion about limits is the demonstration of the proper δ to satisfy the conditions imposed by a given ϵ. Such a demonstration is the proof of the assertion $\lim\limits_{2} (3x - 2) = 4$.

Recall that in the study of geometry proofs were written down in a rather rigid style. It would be helpful to develop a framework or pattern for a limit proof. Let us try. The steps of the proof are given without enclosing them; the explanation of the procedure is enclosed in parenthesis.

$$\text{Step 1} \qquad \left\{ \begin{array}{ll} \text{To prove:} & \lim\limits_{x \longrightarrow 2} (3x - 2) = 4 \\ \text{Proof:} & \text{Let } \epsilon > 0 \\ & \text{Choose } \delta = \square \\ \text{If } 0 < |x - 2| < \delta, & \text{then } |3x - 2 - 4| \end{array} \right.$$

(We have written down the assertion in most succinct form and started our proof by making a statement about ϵ, noting that we must make a choice of δ, and putting down part of the first implication. This procedure, which we label Step 1, is quite mechanical. In Step 3, we shall find a value for δ to be placed in the box thus completing the hypothesis.)

$$\text{Step 2} \qquad \left\{ \begin{array}{l} = |3x - 6| \\ = 3|x - 2| \\ < 3\delta \end{array} \right.$$

(We have changed the form of the expression $|3x - 2 - 4|$ to an equivalent form containing $|x - 2|$. Since our hypothesis asserts "If $|x - 2| < \delta$", we can substitute, thus changing our equality to an inequality. We call this Step 2.)

$$\text{Step 3} \qquad \left\{ \text{Choose } \delta = \frac{\epsilon}{3} \right.$$

(The most important moment in our proof comes when we recognize that we have progressed sufficiently in Step 2 to make a choice of δ. This choice we call Step 3. We can now return to Step 1, fill in the box, and thereby complete the hypothesis.)

$$\text{Step 4} \qquad \left\{ \begin{array}{l} = 3 \cdot \frac{\epsilon}{3} \\ = \epsilon \end{array} \right. \qquad\qquad \text{Q.E.D.}$$

(We have made a choice of δ. This choice we substitute for δ in the last line of Step 2 and simplify. This thereby completes the implications started in Step 1 and we now have a series of steps that show "If $0 < |x - 2| < \delta$, then $|3x - 2 - 4| < \epsilon$". Since this is what we are seeking to demonstrate, our proof is completed. This final substitution and simplification we call Step 4.)

Our conclusion that $\delta = \frac{\epsilon}{3}$ means that the vertical corridor is one-third as wide as the horizontal corridor in Figure 4-3e. We can also see that any other value for δ less than $\frac{\epsilon}{3}$ would define an appropriate closed corridor. Generally we know that if a certain value of δ will work for a given value of ϵ, then any smaller positive value of δ will also work for this ϵ, and in like manner if a certain value of δ will work for a given value of ϵ, then this value of δ will work for any larger value of ϵ.

In the following examples where we use the pattern given above, the steps are labeled, but the explanation is omitted. Each exercise also displays an additional useful maneuver.

Example 1. If $f\colon x \longrightarrow \frac{1}{1 + x^2}$, then prove that $f(x)$ approaches 1 as x tends to 0.

Solution.

To prove: $\lim\limits_{x \to 2} \left(\frac{1}{1 + x^2}\right) = 1$

Step 1 Proof Let $\epsilon > 0$

Choose $\delta = \square$

If $0 < |x - 0| < \delta$, then $\left|\frac{1}{1 + x^2} - 1\right|$

$$= \left|\frac{1 - 1 - x^2}{1 + x^2}\right|$$

Step 2

$$= \frac{x^2}{1 + x^2}$$

$$< x^2$$

$$< \delta^2$$

(We may assert $\frac{x^2}{1 + x^2} < x^2$ since the denominator, $1 + x^2$, is always greater than 1.)

Step 3 $\{$ Choose $\delta = \sqrt{\epsilon}$

(We have not returned to Step 1, filled in the box, and completed our hypothesis. Normally this should be done.)

Step 4 $\{ < \epsilon$ Q.E.D.

If we graph the equation $y = \frac{1}{1 + x^2}$, we obtain Figure 4-3f.

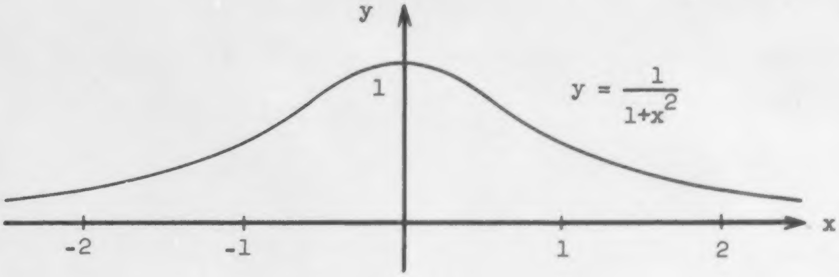


Figure 4-3f.

If we now construct appropriate corridors so that $1 > \epsilon > 0$, we have Figure 4-3g which supports the implications about the size of δ and ϵ arising from the proof.

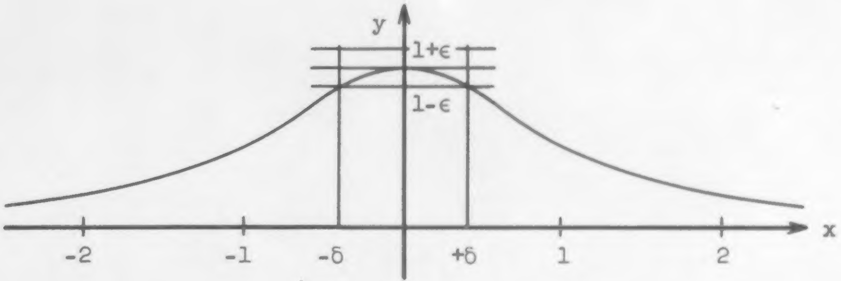


Figure 4-3g.

Remark: The measure of the region under the curve and between the vertical lines $x = 1$ and $x = -1$ is $\frac{\pi}{2}$. We shall come back to this curve when we discuss areas and again when we evaluate π.

Example 2. Let us consider a somewhat more complex problem. Consider the function $f\colon\ x \longrightarrow x^2$.

(1) For values of x close to 3, $f(x)$ is close to 9.

(2) As x approaches 3, $f(x)$ approaches 9.

(3) The difference $|f(x) - 9|$ becomes smaller as the difference $|x - 3|$ is made smaller.

(4) As x approaches 3, the limit $f(x)$ equals 9.

(5) $\lim\limits_{3} f(x) = 9$

All of these statements paraphrase the same fact; all of them are proven in the same manner.

Step 1
$$\left\{\begin{array}{lll} \text{To prove:} & \lim\limits_{x \longrightarrow 3} (x^2) = 9 & \\ \text{Proof:} & \text{Let } \epsilon > 0 & \\ & \text{Choose } \delta = \text{smaller } \square \,,\, \square & \\ \text{If } 0 < |x - 3| < \delta, & \text{then } |x^2 - 9| & \end{array}\right.$$

Step 2
$$\left\{\begin{array}{l} = |(x + 3)(x - 3)| \\ < |(x + 3)\delta| \end{array}\right.$$

(It is not clear as yet what value we should choose for δ. If we add the hypothesis that $0 < |x - 3| < 1$, which is reasonable since we normally are considering x close to 3, then it follows that $x < 4$ and that $|x + 3| < 7$. We add this to the original hypothesis and create an option for the choice of δ. Now if $|x + 3| < 7$, then $|(x + 3)\delta| < 7\delta$.)

Step 3
$$\left\{\begin{array}{l} \text{Choose } \delta = 1 \\ \text{Choose } \delta = \frac{\epsilon}{7} \end{array}\right. \qquad \text{whichever is smaller}$$

Step 4
$$\left\{\begin{array}{l} < |7\frac{\epsilon}{7}| \\ < \epsilon \end{array}\right. \qquad \text{Q.E.D.}$$

Figure 4-3h illustrates the situation when δ is smaller than

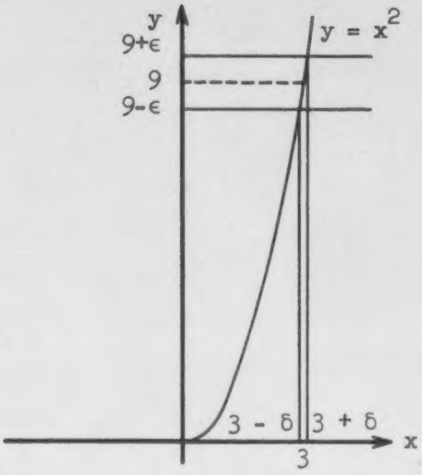


Figure 4-3h.

Example 4. For our final example, let us return to the problem of the derivative mentioned at the beginning of this section. Prove that when $f: x \longrightarrow x^2$, then $f'(x) = 2x$.

Solution.

To prove: $\lim_{x \longrightarrow a} \dfrac{f(x) - f(a)}{x - a} = 2x$ when $f(x) = x^2$

Step 1

Proof: Let $\epsilon > 0$

Choose $\delta = \boxed{\epsilon}$

If $0 < |x - a| < \delta$, then $\left|\dfrac{x^2 - a^2}{x - a} - 2x\right|$

$$= \left|\frac{(x + a)(x - a)}{x - a} - 2x\right|$$

Step 2

$$= |x + a - 2x|$$

$$= |a - x|$$

$$= |x - a|$$

$$< \delta$$

Step 3 $\}$ Choose $\delta = \epsilon$

Step 4 $\} < \epsilon$ Q.E.D.

We conclude this section with a formal definition of a limit of a function.

DEFINITION 4-2. The limit, as x approaches a, of $f(x)$ is L if to each $\epsilon > 0$ there corresponds a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$.

Exercises 4-3

1. Paraphrase each of the following statements in two alternate manners.

   (a) If $f(x) = 2x + 5$, then $f(x)$ approaches 15, as x approaches 5.

   (b) The limit of $\frac{\sin x}{x}$ at $x = 1$ is 1.

   (c) As h gets closer to zero, then $\frac{\sqrt{2 + h} - \sqrt{2}}{h}$ gets closer to $\frac{1}{2\sqrt{2}}$.

   (d) As the difference $|x - a|$ approaches zero, then the quotient $\frac{f(x) - f(a)}{x - a}$ approaches $f'(x)$.

Chapter 5

5-1. Introduction

In Chapter 3, some rules for finding derivatives were developed that are extremely useful. They are not complete, however, and they do not show us how to differentiate the composite functions which arise frequently in applied mathematics. For example, consider the following problem. Two buildings stand on opposite sides of a street 48 feet wide. A ball, dropped out of a window of one building, falls a distance of $16t^2$ feet in t seconds. In the second building an observer is watching from another window at the same height. At what rate is the distance of the ball from the observer increasing two seconds after the ball is dropped?

The distance, s, is found by using a right triangle. See Figure 5-1a.

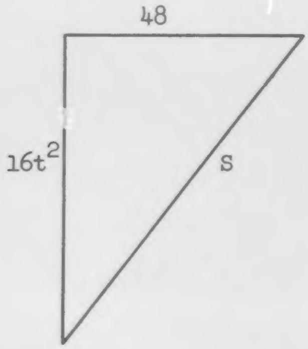


Figure 5-1a.

From the Pythagorean Theorem we have $s = \sqrt{(16t^2)^2 + 48^2}$ and, accordingly, we write $s = f(t)$. Thus the rate of change in the distance when $t = 2$ will be $f'(2)$ Computing $f'(2)$ from the expression

$$\lim_{h \to 0} \frac{f(2 + h) - f(2)}{h}$$

would be quite difficult to say the least. But $f(t)$ can be expressed in the form

$$s = g[h(t)]$$

where $u = h(t) = 256t^4 + 48^2$ and $g(u) = \sqrt{u}$. The derivatives $h'(t)$ and $g'(u)$ are easily determined by rules already learned. Thus we have a question: Can we somehow, through our knowledge of the derivatives of h and g, compute the derivative of the composite of these two functions. Fortunately we have just such a rule called the "Chain Rule" for reasons that will be easily seen and it will prove to be extraordinarily helpful.

5-2. The Chain Rule

It will help in understanding the proof of the following theorem if we recall that, by definition, the derivative is equal to the limit of the difference quotient, or before the limit process is applied, approximately equal to the difference quotient. If the derivative of a function f exists at point b, then

$$\frac{f(y) - f(b)}{y - b} \doteq f'(b) \text{ for } y \text{ close to } b.$$

If we let

$$n(y) = \frac{f(y) - f(b)}{y - b} - f'(b) \text{ for } y \neq b, \; n(y) \longrightarrow 0 \text{ as}$$

$y \longrightarrow b$. If we also define $n(y) = 0$ when $y = b$, it follows that n is continuous at b. Thus

$$f(y) - f(b) = f'(b)(y - b) + n(y)(y - b)$$

holds for both $y \neq b$ and $y = b$. We make use of this equation in the proof of the following theorem.

THEOREM 5-2a. Suppose $g'(a)$ and $f'[g(a)]$ exist. Let $F(x) = f[g(x)]$. Then $F'(a)$ esists and $F'(a) = f'[g(a)] \cdot g'(a)$.

Proof. Let $y = g(x)$ and $b = g(a)$. Thus

$$F(x) - F(a) = f[g(x)] - f[g(a)] = f(y) - f(b).$$

Then [See paragraph above for definition of $n(y)$]

$$F(x) - F(a) = f'(b)(y - b) + n(y)(y - b) \text{ for } y \neq b$$

and $n(y) = 0$ for $y = b$. Dividing by $x - a$, we have

$$\frac{F(x) - F(a)}{x - a} = f'(b)\frac{y - b}{x - a} + n(y)\frac{y - b}{x - a} .$$

Now, according to Theorem 4-7,

$$\lim_{x \to a} \frac{F(x) - F(a)}{x - a} = f'(b) \lim_{x \to a} \frac{y - b}{x - a} + \lim_{x \to a} n(y) \cdot \lim_{x \to a} \frac{y - b}{x - a}$$

provided that the limits exist.

However, $$\lim_{x \to a} \frac{y - b}{x - a} = \lim_{x \to a} \frac{g(x) - g(a)}{x - a} = g'(a).$$

Since $$\lim_{x \to a} g(x) = b \quad \text{and} \quad n \text{ is continuous,}$$

$$\lim_{x \to a} n(y) = \lim_{y \to 0} n(y) = n(b) = 0.$$

Thus

$$F'(a) = \lim_{x \to a} \frac{F(x) - F(a)}{x - a} = f'(b) \cdot g'(a) + 0 \cdot g'(a) = f'[g(a)] \cdot g'(a) .$$

Remark: Although the notation is careless in that the hypothesis is not sufficiently explicit, the Chain Rule can be expressed in alternate notations which are useful when doing problems. For example, if $y = g(u)$ and $u = h(x)$, then $y = f(x)$ where f is a composite function. Accordingly we write

$$f'(x) = g'(u) \cdot h'(x)$$

or $$D_x y = D_u y \cdot D_x u$$

The choice of notation can vary, or course. For example, if $s = g(v)$ and $v = k(t)$, then $s = f(t)$ and

$$f'(a) = g'[k(a)] \cdot k'(a) .$$

We write $$f'(t) = g'(v) \cdot k'(t)$$

or $$D_t s = D_v s \cdot D_t v .$$

Example 1. Find the derivative $D_x y$, given that

$$y = (3x^2 + x + 1)^5$$

Note that this problem can be handled without using the Chain Rule by expanding the trinomial to the fifth power and differentiating term by term.

To do this will unequivocally demonstrate the power of the Chain Rule.

Solution. We observe that we have an equation of the form

$$y = u^5$$

where

$$u = 3x^2 + x + 1 .$$

Adopting the notation used in the note above, we have

$$g(u) = u^5$$

and

$$h(x) = 3x^2 + x + 1 .$$

Hence

$$g'(u) = 5u^4$$

and

$$h'(x) = 6x + 1 .$$

Thus, since $$f'(x) = g'(u) \cdot h'(x) ,$$

we have $$f'(x) = 5u^4(6x + 1)$$

which becomes, when we substitute $3x^2 + x + 1$ for u,

$$D_x y = 5(6x + 1)(3x^2 + x + 1)^4 .$$

Example 2. Find the derivative $f'(x)$, given that

$$f: \ x \longrightarrow f(x) = \sqrt{3x^2 + 1}$$

Solution. Let $y = f(x) = \sqrt{3x^2 + 1}$. If we consider $u = 3x^2 + 1$, then $y = u^{1/2}$. In terms of our earlier notation, this means that $h(x) = 3x^2 + 1$ and $g(u) = u^{1/2}$. Hence we have

$$h'(x) = 6x \quad \text{and} \quad g'(u) = \frac{1}{2}u^{-1/2}$$

Then

$$f'(x) = \frac{1}{2}u^{-1/2} \cdot 6x$$

which becomes, when we substitute for u, rearrange and simplify,

$$f'(x) = \frac{3x}{\sqrt{3x^2 + 1}} .$$

Example 3. Find $f'(x)$, given that

$$f(x) = \left(\frac{x^2}{x + 1}\right)^6$$

Solution. Let $g(u) = u^6$ and $h(x) = \frac{x^2}{x + 1}$. To find $h'(x)$, we must apply our quotient rule. Thus we have

$$h'(x) = \frac{(x + 1)(2x) - x^2(1)}{(x + 1)^2}$$

Also we have

$$g'(u) = 6u^5$$

Applying the Chain Rule, we obtain

$$f'(x) = 6u^5\frac{(x + 1)(2x) - (x^2)}{(x + 1)^2}$$

Substituting $h(x)$ for u and simplifying, we have

$$f'(x) = \frac{6(x^{12} + 2x^{11})}{(x + 1)^7}$$

Now we are ready to complete the problem that was used to introduce Section 5-1.

Example 4. Let $f(t) = \sqrt{(16t^2)^2 + 48^2}$, find $f'(2)$.

Solution. Consider $u = 256x^4 + 2304$ and $y = \sqrt{u}$. Then $D_x u = 1024x^3$ and $D_u y = \frac{1}{2}u^{-1/2}$. Hence $D_x y = (\frac{1}{2}u^{-1/2})(1024x^3) = \frac{512x^3}{\sqrt{256x^4 + 2304}}$.

Substituting $x = 2$, we have $f'(2) = 51.2$ feet/second.

More of the power of the Chain Rule (and the appropriateness of the name) will be revealed in the following extension. Consider $D_x f[g\{h(x)\}]$. If we let $s(x) = g\{h(x)\}$, then, according to Theorem 5-1, we have

$$D_x f[g\{h(x)\}] = D_x f[s(x)] = f'[s(x)] \cdot s'(x) .$$

But $s(x)$ is a composite function. Accordingly

$$s'(x) = g'\{h(x)\} \cdot h'(x) .$$

By substitution we have

$$D_x f[g\{h(x)\}] = f'[g\{h(x)\}] \cdot g'\{h(x)\} \cdot h'(x) .$$

If we have three defining equations of the form $y = f(v)$, $v = h(u)$, and $u = g(x)$, then we write

$$D_x y = D_v y \cdot D_u v \cdot D_x u .$$

Example 5. Find $D_x y$, if $y = \sqrt{\left(\frac{x^2 + 1}{x}\right)^3}$.

Solution. Let $y = \sqrt{u}$, $u = v^3$, and $v = \frac{x^2 + 1}{x}$. Then $D_u y = \frac{1}{2}u^{-1/2}$, $D_v u = 3v^2$, and $D_x v = \frac{x(2x) - (x^2 + 1)}{x^2}$. By the extended Chain Rule, we have

$$D_x y = \frac{1}{2}(v^3)^{-1/2} \cdot 3\left(\frac{x^2 + 1}{x}\right)^2 \cdot \frac{x^2 - 1}{x^2}$$

$$= \frac{3\left(\frac{x^2 + 1}{x}\right)^2\left(\frac{x^2 - 1}{x^2}\right)}{2\sqrt{\left(\frac{x^2 + 1}{x}\right)^3}} = \frac{3(x^2 - 1)(x^2 + 1)^2}{2x^3\sqrt{\left(\frac{x^2 + 1}{x}\right)^3}}$$

Exercises 5-2

1. Find $D_x y$ in each of the following

   (a) $y = (3x^2 + 1)^2$

   (b) $y = \sqrt{4x - 1}$

TEACHER'S COMMENTARY

Chapter 3

THE DERIVATIVE

3-1. A Problem (Introduction)

One of the two basic ideas of the elementary calculus is "derivative." It is easy to appreciate this idea intuitively and know why it is useful before formulating it precisely. Here we consider this idea as it arises in the solution of a specific problem.

The purpose of this chapter is to polish and extend the concept of the difference quotient. We finally arrived at a meaningful definition and useful formulas for the derivative. The difference quotient is a ratio whose denominator is the difference between two elements of a domain and whose numerator is the difference between the corresponding values of the range. Use of the word *ratio* is well advised since one interpretation of a difference quotient will be as a *rate* of change.

As indicated in the preface, the text insofar as possible is problem motivated. Here, as we begin the study of the derivative, a familiar problem is explored at length. By doing this we seek to give the student an overall point of view and to instill an appreciation of the approach a mathematician takes when attacking a problem.

This introduction should be read and discussed briefly. The more the appetite is whetted the better, but do not bog down in explanation. The next several chapters are devoted to this task.

3-2. Numerical Computation of Slope

It is intended that the following three important ideas be emphasized throughout this computational exercitation: (1) approximation of slope, (2) linear approximation, and (3) method of approximation.

(1) Unless the student has studied the section on tangents to graphs of polynomials in SMSG *Elementary Functions* or the equivalent,

his experience dictates that slope is a number describing the constant rise over run of a straight line. Here he must feel comfortable with the assumption that the direction of a curve at a point can be described by a straight line approximation.

(2) Henceforward "best linear approximation" and "tangent line" may be used interchangeably.

(3) We take points successively closer to and on either side of the point under consideration. A lengthy discussion of a possible inflection point where the tangent line "passes through" the curve is intentionally postponed. This topic will be handled thoroughly in a treatment of concavity with regard to curve tracing (Section 6-2) after a study of higher order derivatives (Section 5-6). In the example of this section we can safely use the "squeeze" method of successive approximations since it tangibly suggests the limit of the difference quotient. In "squeezing" the slope of the tangent between two numbers very close together, we allude to the intuitively appealing "flyswatter" (or "sandwich") theorem.

The "flyswatter" process (used in this text as early as Section 1-1) has been employed significantly by mathematicians over the years. Because of its import in the development of this text, we submit here Theorem 4-4g (the "flyswatter" theorem), the proof of which appears in Chapter 4 of the student text.

THEOREM 4-4g. Suppose that for some $s > 0$, $g(u) \le f(u) \le h(u)$ for $0 < |u - a| < s$. Further suppose that $\lim_{u \to a} g(u) = L$ and $\lim_{u \to a} h(u) = L$. Then we can conclude $\lim_{u \to a} f(u) = L$.

Before leaving this computational section (the students concurring wholeheartedly with the authors' second sentence in Section 3-2), a small positive statement about "ϵ" is advisable. Although the clause "where ϵ represents any small positive number arbitrarily chosen" appears only once in Section 3-1, the clause is implied each time that ϵ is used. Whereas no editor would stand for monotonous repetition of this clause in a text, the teacher might find that any calories expended in this pursuit will be B. T. U.'s well spent upon arrival at the marriage of ϵ and δ in Chapter 4. We propose to have the student "keep his eye on the epsilon," that is, be concerned with a change in $f(x)$ without worrying initially about a change in x.

3-3. The Slope Function for $y = x^2$

The following three ideas should be emphasized in this section: (1) a rule can be formulated for giving the slope of the function x^2 and of the general type x^n, (2) the rule for slope itself constitutes a new function, and (3) the functions with which we are currently dealing have smooth graphs with no breaks.

(1) The so called "power rule" which is derived here is treated more completely with a broader basis for application in Section 5-2.

(2) It is important for the work that follows (immediately and later in antidifferentiation) that the student think of the slope of a given function as a function itself. It would be time well spent if the student were required to graph the slope functions of many early functions he examines. The graph of the slope function of a given function can be used as a summary interpretation of the rate of change of the given function.

(3) In the work of this section the student is reminded that the curves under consideration have no "jumps or breaks." To suggest reasons for these stipulations the teacher may wish to submit some counter examples where the slope cannot be found at a certain point. For example the functions $F: x \longrightarrow \frac{1}{x}$ and $G: x \longrightarrow \frac{x^3}{x}$ are discontinuous at $x = 0$ and $H: x \longrightarrow \frac{x^2 - 1}{x - 1}$ has a "hole" at $x = 1$. Consequently no slope exists at a point which doesn't

exist. Illustrative of curves that have sharp "corners" are graphs of the functions $f\colon x \longrightarrow |x|$, $g\colon x \longrightarrow x^2s$, and $h\colon x \longrightarrow |4 - x^2|$. The slope is not defined at $(0,0)$ for f and g nor at $(-2,0)$ and $(2,0)$ for h.

Before going on we propose an exercise which the student can handle at this point. Having been led through this problem he will have "discovered" the slope function of $f\colon x \longrightarrow |x|$, and may feel content that he has done some amateur analysis. After first requiring him to sketch separate graphs for $f\colon x \longrightarrow |x|$ and $\phi\colon x \longrightarrow \frac{|x|}{x}$, pose the following questions.

(a) What is the slope of the graph of $f\colon x \longrightarrow |x|$ for $x > 0$? for $x < 0$?

(b) What is the range of $\phi\colon x \longrightarrow \frac{|x|}{x}$ for $x > 0$? for $x < 0$?

(c) Employing your conclusions from (a), graph the slope function f', where $f\colon x \longrightarrow |x|$ for $x > 0$ and $x < 0$.

(d) Compare the graph of the slope function f' with the graph of the function $\phi\colon x \longrightarrow \frac{|x|}{x}$. What does geometric evidence suggest regarding the identity of the slope function of $|x|$?

3-4. The Velocity Function

In this section we propose to build upon the student's previously acquired notions regarding velocity and speed. His experience, however, probably does not dictate a distinction between velocity and speed. Speed is the absolute value of velocity.

Our primary aim begin to relate velocity and the slope function, we submit here a supplementary section TC 3-4. This addendum may be used to lead the student through some restricted analysis of the velocity function within the framework of an elementary problem. Besides the basic problem, there are two sets of related exercises. Exercises TC 3-4(a) may be employed to motivate some of the ideas in Section 3-4, while exercises TC 3-4(b) may well be postponed until Section 3-5 has been studied.

TC 3-4(a). An Elementary Projectile Problem

Let us assume that a pellet is projected straight up and after a while comes straight down via the same vertical path to the place on the ground from which it was launched. After t seconds the pellet is s feet above the ground. Some ordered pairs of the form (t,s) are given in the following table.

Table TC 3-4(a)

t	0	1	2	3	4	5	6	7	8	9	10
s	0	144	256	336	384	400	384	336			0

We shall intentionally avoid certain physical considerations such as air resistance. Moreover, we shall deal with "nice, round" numbers rather than quantities measured to some prescribed degree of accuracy which might arise from the data of an actual projectile problem in engineering. In the exercises that follow our attention will be directed to certain mathematical principles, particularly basic concepts of differentiation.

TC 3-4(a) Exercises

1. Interpolate from the data given to determine the height of the projectile after eight and nine seconds respectively. (Guess, using symmetry as your guide.) Does extrapolation to find values of s for $t = -1$ or $t = 11$ make sense on physical grounds? After how many seconds does the projectile appear to have reached its maximum height? What seems to be the maximum height?

2. Does s appear to be a function of t? What is probably the domain of t? What seems to be the range of s?

3. If we were to plot a graph of $s = f(t)$,

   (a) is it plausible on physical grounds to restrict our graph to the first quadrant?

   (b) does the data suggest that the scale on the s-axis (vertical) should be the same as the scale on the t-axis (horizontal)?

4. Keeping in mind your responses to exercise 3, plot the ordered pairs (t,s) from the table. Connect the points with a smooth curve. What is the name of the function suggested by the graph? On physical grounds is it feasible that there would be a real value of s for every real number assigned to t over the interval $0 \le t \le 10$? Were we probably justified in connecting the points?

5. Assuming that the equation $s = f(t) = at^2 + bt + c$ was used to develop Table 3-3, find values for constants a, b, and c.

6. Sketch the graph of $\{(t,s): s = 160t - 16t^2\} \cap \{(t,s): 0 \le t \le 10\}$. Using a more carefully plotted graph of the above set, connect the point where $t = 1$ with the point where $t = 2$ with a secant line. What is the slope of this secant? Draw tangents to the curve at $t = 1$ and $t = 2$, and estimate their slopes from your graph.

7. If the units of s are feet and the units of t are seconds, what are the units of slope? What word is commonly associated with this ratio of units? What would you guess are the physical interpretations of positive, zero, and negative values of this ratio?

8. Draw the graph of $v = 160 - 32t$ over the interval $0 \le t \le 10$. Compare the values of v for $t = 1$ and $t = 2$ respectively with your estimates for the slopes of tangents to the graph of $s = 160t - 16t^2$ in exercise 6.

9. Average the values of v for $t = 1$ and $t = 2$ and compare this average with the slope of the secant connecting the points where $t = 1$ and $t = 2$ in exercise 6.

10. If the units of v are ft/sec and the units of t are seconds, what are the units of the slope of the line $v = 160 - 32t$? What word from physics is commonly associated with this ratio of units? Does the minus sign along with the particular numerical value of this slope have any special connotation from your experience?

TC 3-4(b). More About the Elementary Projectile Problem

The preceding exercises were intended to motivate a consideration of certain relationships between a function, its graph, its rate of change, and the graph of its rate of change within the restricted framework of an elementary physical problem. Let us now consider the same problem in more depth.

The projectile is fired vertically from the ground so that after t seconds its distance s from the ground is given by the equation

$$s = f(t) = 160t - 16t^2.$$

Since the height of the projectile depends upon the number of seconds, we shall regard distance as a function of time. Since the range of this function is restricted to $s \geq 0$ because of physical condition, the domain is $0 \leq t \leq 10$.

The physical basis of the problem should convince us that we can graph the parabola $s = 160t - 16t^2$ in the first quadrant as a continuous function, since for each "split-second" of time there is a corresponding height. However, we should keep in mind that while the graph of $s = f(t)$ is parabolic, the projectile itself ascends and descends in a straight vertical path. We should not confuse the path of the projectile with the graph of its distance function.

A graphical representation of average velocity from t_1 to t_2 seconds is the slope of a secant line connecting the point (t_1,s_1) with the point (t_2,s_2) on the graph of $s = f(t)$. Instantaneous velocity, or the velocity at a particular time t, is equal to the slope of a tangent to the graph of $s = f(t)$ at the point (t_1,s_1). Thus velocity is interpreted as rise/run of a straight line with units ft/sec.

For this example it is true that the numerical average of the instantaneous velocities at t_1 and t_2 is the same as the average velocity from t_1 to t_2. However, we should not be too quick to generalize to other velocity problems.[1]

From an examination of the tangent s to the parabolic graph of $s = f(t)$ we observe that at different points the tangents have different slopes. The fact that the slope of f(t) is dependent upon the value of t is concrete evidence that the slope is itself a function of t. The physical counterpart of this inference for our problem is that velocity, as well as distance, is a function of time. On the basis of this intuitive argument we shall define instantaneous velocity as the slope function or derivative, f'(t), of the distance function s = f(t).

---

[1] "Many a beautiful theory has been 'shot down' by an ugly fact." For example, a motorist traveling from Candlestick Park to Dodger Stadium at an average rate of 60 miles per hour and returning over the same route at an average rate of 30 miles per hour does not average 45 miles per hour for the entire journey.

$$\boxed{v = s' = f'(t)}$$

Let us now find the derivative of the function $f(t) = 160t - 16t^2$, consulting the graph of s = f(t) for a geometric interpretation of our procedure. See Figure TC 3-3(b). We shall consider the limit as h approaches zero here as an alternate to the difference quotient limit technique.

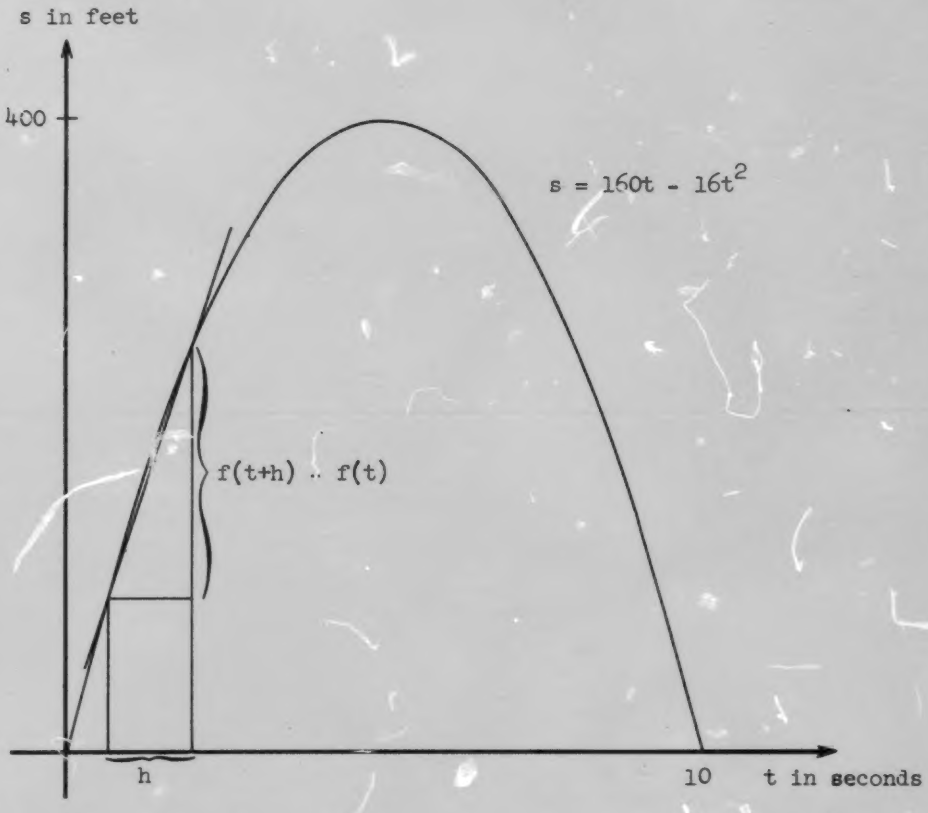


Figure TC 3-4(b)

$$f(t) = 160t - 16t^2$$

$$f(t + h) = 160(t + h) - 16(t + h)^2$$

$$= 160t + 160h - 16t^2 - 32ht - 16h^2$$

$$f(t + h) - f(t) = 160h - 32ht - 16h^2$$

$$\frac{f(t + h) - f(t)}{h} = 160 - 32t - 16h$$

For h very small, $160 - 32t - 16h \approx 160 - 32t$

Finally, $\lim\limits_{h \to 0} 160 - 32t - 16h = 160 - 32t$

Thus, $\lim\limits_{h \to 0} \frac{f(t + h) - f(t)}{h} = 160 - 32t$

By definition $\lim\limits_{h \to 0} \frac{f(t + h) - f(t)}{h} = f'(t)$

Therefore, $f'(t) = 160 - 32t$

Consequently, $v = 160 - 32t$

Exercises TC 3-4(b)

1. Graph the function $v = 160 - 32t$ over the interval $0 \leq t \leq 10$.
2. (a) What is the v-intercept of the function $v = 160 - 32t$?
   (b) What is the slope of the tangent to the graph of $s = 160t - 16t^2$ at the point where $t = 0$?
   (c) What is the initial velocity, v_0, of our projectile?
3. What is the average velocity of the projectile between:
   (a) The first and second seconds of flight?
   (b) The eighth and ninth seconds of flight?
4. (a) What is v when $t = 10$?
   (b) What is the impact velocity of the projectile?
5. After six seconds of flight:
   (a) What is the height of the projectile above the ground?
   (b) What is the total distance traveled by the projectile?
   (c) What is the velocity of the projectile?
   (d) Is the projectile ascending or descending?

6. Explain the difference in sign between your two answers for exercise 3 and between the initial velocity and impact velocity.

7. (a) What is the t-intercept of the function $v = 160 - 32t$?
   (b) For what value of t is the tangent line to the graph of $s = 160t - 16t^2$ horizontal?
   (c) After how many seconds of flight does the projectile appear to "stop for an instant" before descending?
   (d) How many seconds have expired when the projectile reaches its maximum height?
   (e) What is the maximum height attained by the projectile?

8. (a) If $v = s' = f'(t) = 160 - 32t$ and $\alpha = v' = f''(t)$, find α.
   (b) What is the "rate of change of the rate of change" (acceleration or deceleration) of the distance function $s = f(t)$?
   (c) What kind of a function is α? Sketch it.
   (d) Taking into account the sign and numerical value of α together with the physical properties of this problem, what is your interpretation of α?

9. (a) Shade the area given by $\{(t,\alpha): -32 \le \alpha \le 0\} \cap \{(t,\alpha): 0 \le t \le 5\}$.
   (b) How many square units of area have been shaded?
   (c) Is this number of square units numerically the same as your answers for exercise 2?

10. (a) Shade the area given by $\{(t,v): 0 \le v \le 160 - 32t\} \cap \{(t,v): t \ge 0\}$.
    (b) How many square units of area have been shaded?
    (c) Is this number of square units numerically the same as your last answer for exercise seven?

After the student has studied the chapter on the definite integral, he might be asked to return to this set of exercises and explain his answers for part c of exercises 9 and 10.

3-5. The Derivative

We have discussed two interpretations of the derivative, slope and velocity. We saw that in every case a slope function or a velocity function was derived from some initially considered function. Soon we shall wish to point out that the process of finding the derivative is called differentiation, since it involves taking the differences $f(x) - f(a)$ and $x - a$. In Chapter 1 we made the following statement. "Employing the notion of differences to reach a useful definition of the word 'limit' proves to be very fruitful, as we shall see." Indeed we shall see this very shortly as we define limit formally in Chapter 4. But at this point we are in a position to make the extension from the difference quotient with its geometric counterpart to the definition of derivative. Because this transition is so important we outline here (TC 3-5, The Derivative) a supplement to the content of Section 3-5 which incorporates the principal ideas of the entire chapter.

TC 3-5. The Derivative

a. Geometric Interpretation

Let $P = \big(a, f(a)\big)$ and $Q = \big(x, f(x)\big)$ be two points on the graph of $y = f(x)$. The slope of PQ is given by

$$\frac{f(x) - f(a)}{x - a} = F(x)$$

If $F(x)$ approaches a limit as $x \longrightarrow a$, we shall call this limit the derivative of f at a and denote it by the symbol $f'(a)$. Geometrically, $f'(a)$ represents the slope of the tangent to the graph drawn at the point P.

b. A Working Definition

The function f is said to have the derivative, $f'(a)$, at a, if

$$\frac{f(x) - f(a)}{x - a} \text{ approaches } f'(a) \text{ as } x \text{ approaches } a.$$

c. A Useful Form

We have written $\frac{f(x) - f(a)}{x - a} = F(x)$ where $x \neq a$. (1)

By definition $f'(a) = \lim\limits_{x \longrightarrow a} F(x)$

If we supplement (1) by the agreement that $F(a) = f'(a)$, F is defined for all values of x in the domain of f and is continuous at $x = a$. We may write

$$f(x) = f(a) + (x - a)F(x) \tag{2}$$

Since equation (2) holds for $x = a$ as well as $x \neq a$, we can employ it in more situations than equation (1).

d. Error of Linear Approximation

Let us compare equation (2), which represents the curve, with the linear equation

$$g(x) = f(a) + f'(a)(x - a) \tag{3}$$

which represents the tangent to the curve at $x = a$. The two equations differ in only one respect. In equation (2) the coefficient of $(x - a)$ is $F(x)$, whereas in equation (3) it is $f'(a)$. Subtracting (3) from (2) we obtain

$$f(x) - g(x) = [F(x) - f'(a)](x - a).$$

Let us call this difference $\eta(x)$, where $\eta(x)$ represents the error committed in using linear g to approximate the f curve. That is,

$$\eta(x) = [F(x) - f'(a)](x - a). \tag{4}$$

Dividing both sides of (4) by $(x - a)$ we have,

$$\frac{\eta(x)}{x - a} = F(x) - f'(a), \quad x \neq a.$$

Since $\frac{(x)}{x - a} = F(x) - f'(a) \longrightarrow 0$ as $x \longrightarrow a$, we may choose x sufficiently close to a so that the error of linear approximation is an arbitrarily small fraction of $x - a$.

e. Notation for the Derivative

With an obvious preference for the definition of the derivative of f at a as the limit of a difference quotient, namely

$$\lim_{x \longrightarrow a} \frac{f(x) - f(a)}{x - a},$$

the authors feel that it is important for the student to recognize and become facile with other forms. For example, if we use variable notation where $y = f(x)$, the following should have essentially the same meaning

$$\lim_{h \longrightarrow 0} \frac{f(x + h) - f(x)}{h}$$

$$f'(x)$$

$$D_x y$$

$$\lim_{\Delta x \longrightarrow 0} \frac{f(x) + \Delta x) - f(x)}{\Delta x}$$

$$\lim_{\Delta x \longrightarrow 0} \frac{\Delta y}{\Delta x}$$

$$y'$$

The derivative function may be noted as f' or D_x.

Notice that we have omitted the quotient from $\frac{dy}{dx}$ for now. This omission is intentional. We feel that it is advisable to avoid this quotient form here because we want the students' attention to be focused upon the fact that the word "derivative" is used at once as a function and a limit. To appreciate the significance of the $\frac{dy}{dx}$ notation the student must be able to understand differentials. Differentials are postponed until Section 6-6, where the increments Δy and Δx are discussed in a sophisticated context of approximations.

PART B

The Basis for the Decision to Recommend an SMSG Calculus Text

The first task of our writing team was to examine existing calculus books to determine whether there already existed an adequate number of calculus texts meeting the desired objective. This objective is to provide for high school students with ninth- and eleventh-grade SMSG background a one-year calclulus course which would enable them to receive credit for one year of calculus upon entering college. Although there are many excellent calculus books currently in print, it was our feeling that none of the books is entirely satisfactory for the expressed purpose. It has therefore been decided to recommend that an SMSG calculus text be written.

The principal ways in which the considered texts failed to be suitable are listed below:

(A) failed to build on SMSG background;
(B) over-concentration on theory;
(C) too weak on theory;
(D) too formal a style of presentation;
(E) too sophisticated in language;
(F) too unconventional approach to the subject;
(G) too concise;
(H) insufficient motivation of ideas;
(I) unconventional notation;
(J) too much material for a one-year course.

Though no book was rejected on the basis of (J) alone, this reason is more important than it may at first appear. A book designed for a two-year course often requires a good deal of "skipping around" when being adapted for a one-year couse. Also, it is sometimes the case that theorems are stated and proved in the early part of the book that have their only application in the latter part.

I would like to add an impression of my own to the above list. Although it is no doubt the case that students taking a calculus course in high school are very highly selected, they are still high school students in a high school environment, and I do not feel that we can always obtain from them the same down-to-business attitude that we can demand of university students. For this reason, I feel that some otherwise satisfactory texts will fail to qualify on the grounds of being too colorless. I feel that we have to continually recapture the student's interest by motivating our ideas with geometrical, physical, or realistic problems.

Prefatory Remarks

It having been decided that an SMSG calculus text should be written, the next problem was to decide what should go in it and how it should be organized. The group arrived at the "boundary conditions" enumerated below. Some of these conditions will be explained and elaborated upon in the following pages of these remarks.

(1) The text should be designed to cover that material normally encountered in a first-year college calculus course.

(2) It should prepare the student to take and pass an advanced-placement exam in calculus.

(3) It should be confined to the calculus of functions of one variable, not including infinite series.

(4) It should include SMSG mathematics through Intermediate Mathematics (Elementary Functions not to be a prerequisite).

(5) Analytic geometry is not to be included except for a bried review of those topics necessary to calculus.

(6) The course is to be oriented toward application and problem solving rather than toward theory.

(7) It should include an introduction which will explain what calculus is about.

(8) The concept of the derivative is to precede and to be used to motivate the concept of limit.

(9) A fairly thorough treatment of ϵ's and δ's is to be given.

(10) All concepts and theorems are, whenever possible, to be motivated by natural geometrical or physical problems.

(11) The logarithm is to be defined as $\int_1^x \frac{1}{t}\,dt$ and the exponential as the inverse function of the logarithm.

(12) The exponential and logarithm functions are to be obtained as soon as possible for use in applications.

(13) The integral is to be motivated by area and to be arrived at through upper and lower estimates.

(14) Only those applications of the derivative which are necessary to bring home its meaning and importance should precede the introduction of the integral.

(15) The definite integral should precede the indefinite integral.

(16) The definition of the definite integral as the limit of a sum is to be highly stressed before the FTC is taken up.

(17) All notation involving variables is to be deferred until the subject of differentials is taken up.

(18) Numerical methods and the idea of arbitrarily close approximation is to be emphasized in the text.

These boundary conditions to a large extent determine the subject matter and the style of the text and, in fact, almost uniquely determine the order of the material in the first eight chapters. We will now explain and comment on some of those boundary conditions.

(1), (2), and (3). The standard first-year calculus course consists of the calculus of functions of one variable, not including infinite series. This is the material tested in the standard advanced-placement tests.

(4), (5), and (11). The outstanding features of the SMSG background mentioned in (4) are the following. It includes a considerable amount of analytical geometry, much more, in fact, than is strictly necessary for calculus, so that only the briefest review is necessary. In fact Cartesian coordinates in the plane are so firmly ingrained that it is not necessary to say anything about them; we just use them. Number systems have been discussed at great length, including a little work on the real numbers. Field and order axioms have been drummed in year after year. There has been considerable work on inequalities and absolute value so that it is only necessary to review these topics. In the Intermediate Math text the logarithm was introduced by means of area--thus explaining condition (11). Another asset of the SMSG background is a much more mature way of looking at mathematics than that engendered by the conventional curriculum.

(6) and (10). Condition (6) would seem at first glance to be entirely at variance with (9), and it is indeed rather difficult to reconcile them. Perhaps it will help to explain (6) as we understand it. We feel that the idea of mathematical proof is fundamental to any mathematics course and cannot be dispensed with. We intend to clearly state just which unproved propositions we are assuming in the text, and then we will use logical reasoning to obtain other theorems. We will not be absolute purists in this matter, however. For example, in defining the trigonometric functions, we speak of measuring a distance x around the unit circle. Here we are using heuristic in tacitly assuming that this makes sense. In obtaining the inequality $\sin x < x < \tan x$, we again use heuristic in assuming the obvious but unproved properties of area. I also feel that the

existence of the derivative of the inverse function should be left to heuristic. Also in applications such as volume of solid of revolution and volume of a solid of known cross section, we will use intuitive ideas of volume. But in the main, the course will be firmly founded on our axioms which are the field, order, and several completeness axioms for real numbers and the maximum and intermediate value properties and the existence of the integral for continuous functions. If an ϵ-δ treatment of limits were to be omitted, we would add half a dozen limit theorems to this list.

I do not feel that the exceptions noted above constitute the primary meaning of condition (6). Its principal meaning to me is that the theory should be unobtrusive. Most of our theorems and proofs will not be labeled as such but will be casually presented in the text. The student will not be asked to commit these theorems or their proofs to memory or to refer to them by number. Furthermore the theory should be motivated by natural problems and grow out of intuitive ideas.

We have not been able to exemplify these principals in the student materials to the degree we would like, being primarily involved as we were with the problem of laying out the theoretical groundwork in accord with the above conditions. Some points which do exemplify these principals are:

(a) the use of tangent lines to motivate the derivative, which in turn motivates limits;

(b) the use of a specific problem to motivate the chain rule;

(c) the extensive use of area to motivate the integral.

We are sorry that we have not had any time to give to problems which probably constitute a good half of the value of any calculus course.

The remainder of this report consists of a table of contents for the proposed text, explanatory notes for Chapters 1 through 8, and student materials in various stages of completion for some of these chapters. The explanatory notes are intended primarily for future writing teams and secondarily for the SMSG Advisory Board. The correct order of topics after Chapter 8 has not been worked out.

At this point we should like to express our thanks to Peter Lax, William Lister, J. L. Kelley, Annalee Lax, and Leon Cohen, who have made many valuable suggestions.

Table of Contents
for the Proposed SMSG Calculus Text

Chapter 1. Introduction

1.1 The Concept of the Derivative

1.2 The Concept of the Integral

1.3 Problems Solvable Using These Concepts

Chapter 2. Review

2.1 The Real Number System and its Properties

Field and order properties. Completeness property. Real numbers as points on a line.

2.2 Inequalities and Absolute Value

Geometrical interpretation. Intervals. Triangle inequality.

2.3 Functions

The function concept, unique determination. Ways of representing functions.

2.4 Some Types of Functions

Monotone functions. Inverse functions.

2.5 Special Functions

Linear functions. Trigonometric, inverse trigonometric functions and their graphs. Trigonometric identities and inequalities.

Chapter 3. The Derivative

3.1 Definition of the Derivative

Geometrical meaning of derivative. Intuitive idea of a limit. The derived function.

3.2 Derivatives of Some Power and Root Functions

Derivatives obtained from definition using the intuitive idea of limits.

3.3 Velocity

3.4 Derivatives of the Sine and Cosine

The limit as $x \to 0$ of $\frac{\sin x}{x}$. Its use in differentiating $\sin x$.

Chapter 4. Limits

4.1 Definition of Limit

Refinement of the intuitive idea of limit.

4.2 Limit Techniques

Application of definition to simple examples.

4.3 Some Limit Theorems

Sum, product, quotient, squeeze.

4.4 Continuity and Further Limit Theorems

Definition of continuity. Composition theorem. An important corollary.

4.5 Continuity on an Interval

Maximum and intermediate value theorems.

Chapter 5. Theory and Applications of the Derivative

5.1 Rules for Differentiation

Sum, product, quotient rules. Derivatives of polynomials, trigonometric functions.

5.2 Best Linear Approximation

$f(x) - f(a) = f'(a)(x - a) + \mu(x)(x - a)$ where $\mu(x) \to \mu(a) = 0$ as $x \to a$.

5.3 The Chain Rule

The chain rule. Derivatives of inverse functions. Derivatives of inverse trigonometric functions.

5.4 Maxima and Minima

Physical and geometrical applications.

5.5 The Law of the Mean

Rolle's theorem. Law of the mean. Constancy of a function whose derivative is zero on an interval. Monotone functions.

5.6 Curve Plotting

Maximum and minimum intervals of monotoneity. Zeros, asymptotes, extent.

Chapter 6. Area and Integral

6.1 The Intuitive Concept of Area

Approximation of area under curve $f(x) = x^2$.

6.2 Sigma Notation

Various summation formulas.

6.3 Computation of Areas

Area under curve $f(x) = x^2$ and other curves.
Upper and lower sums.

6.4 Area under Curve $f(x) = \cos x$.

6.5 The Definite Integral

Definition, existence, uniqueness for monotone functions, for continuous functions.

6.6 Riemann Sums

6.7 The Mean-Value Theorem and Other Theorems

The consequence of $f(x) \leq g(x)$.

Chapter 7. The Fundamental Theorem of Calculus

7.1 The Fundamental Theorem (two terms).

Integral of the derivative and derivative of the integral.

7.2 The Indefinite Integral

Antiderivative. Some integration formulas.

7.3 Some Properties of the Integral

Integral of the sum. Substitution theorem.

7.4 Tabulating the Sine and Cosine

Chapter 8. The Logarithmic and Exponential Functions

8.1 The Function $L(x)$

8.2 The Function $E(x)$

8.3 Applications

Chapter 9. Application of the Integral

Area between two curves. Volume of revolution (two methods). Work. Fluid pressure. Solids of known cross section.

Chapter 10. Further Properties of the Derivative

Differentials. Approximation. Variables. Other notations for the derivative. Higher derivatives. Acceleration. Concavity. Further curve plotting. Implicit differentiation. Some simple differential equations.

Chapter 11. Techniques of Integration

Substitution, integration by parts, trig substitution. Powers of trig functions. sin ax cos bx. Partial fractions. Applications.

Chapter 12. Approximation and Numerical Methods

Newton's method. Taylor's formula. Trapezoid rule. Simpson's rule. Error. Arc length. Area of surface of revolution.

Chapter 13. Parametric Equations and Polar Coordinates

Chapter 14. An Album of Advanced Problems

Chapter 15. A Look Ahead

An indication of the extension of the ideas of calculus into various fields of analysis.

Explanatory Remarks for Chapter 1--
Introduction

What does a student want to know when embarking on a new course of study? I feel that he most of all wants to know what the subject is about. My own experience has been that, when commencing to read a mathematical work in a field with which I am unfamiliar, I find it most difficult to follow the preliminary theorems and definitions unless there is an introduction explaining the central ideas of the theory and telling where the theory is going and what it can do.

Many calculus books commence with material on inequalities, absolute value, real numbers, functions, limits before the derivative is taken up. It is difficult for the student to see where all this is leading. He does not see, when he is studying this background material, that there is a unifying subject, the derivative, that will tie it all togethe . He has no idea, in these early stages, or what the subject is all about.

My feeling is that such a presentation as described above is rather like presenting a body of evidence in court without having said what the case is that is being tried. The auditor listening to this evidence will find it impossible to remember the evidence or to make any sense out of it or to evaluate it in any way without the case to provide a framework in which to store, arrange, and systematize it in his mind.

I am particularly dubious of the value of a historical approach in the introduction for the reason that I feel that a student can have no appreciation for the history of a subject which he knows nothing about. The proper time to discuss the history of an idea is after the reader knows what the idea is. It is proposed to avoid these pitfalls by explaining at the outset what calculus is about by solving two very natural problems in which the derivative and the integral arise. It is hoped that such a "preview" will provide the student with a frame of reference in which to fit his subsequent knowledge. As each new idea or theorem is presented the student will be able to see how it is related to the central problem. There is the further advantage that tangent lines and areas are natural and interesting problems, while limits are unnatural and inequalities and absolute value seem by and large rather dull to most students.

Another value of such a beginning is that it exhibits the difference between calculus and algebra. The student has no doubt heard that calculus is a "different kind" of mathematics, and he wants to know how it is different. Our approach to the subject will answer this question at the outset. Still a further value of this procedure is that it emphasizes the essential simplicity and naturalness of the basic concepts of derivative and integral. Later on when the going gets rough, the student will be able to recall that he had no difficulty in understanding the basic concepts at the outset, so that things cannot be so bad after all.

Two versions or "forms" of such an introduction are included. The first form is extremely polished and could hardly fail to be extremely interesting to the student. The second form, which is rather rough by comparison, is included only because it embodies the feature of exhibiting the concepts of derivative and integral in solitary splendor without the extraneous idea of optimization. We want to be sure that the student knows what the concepts of derivative and integral are and does not mix them up with something else.

Explanatory Remarks for Chapter 2--
Review

The only student materials submitted for this chapter consist of a section introducing and clarifying the concept of a function.

The following remarks have the purpose of explaining what material must be included in this chapter according to our overall plan.

2.1 The Real Number System and its Properties

First we review the properties of real numbers as presented in the SMSG Intermediate Mathematics text. These are

Field Axioms

Order Axioms.

(There is some difference of opinion as to whether some remark should be made here to indicate that the rational numbers already have these properties. The property that serves to distinguish the real numbers from the rationals is found in the following axiom which holds in the reals but not in the rationals.)

Completeness Axiom

This axiom will not actually be called by this name. There are three forms which will be useful. It might be stated that they are equivalent, but it will certainly not be proved. The three forms are

(a) every real number has a unique representation as an infinite decimal.

(b) every real number has a unique representation as a point on the number line.

(c) the separation axiom.

The representation of real numbers as infinite decimals was treated in Intermediate Mathematics and should be referred to here in order to avoid giving the student the impression that we are abandoning the old concept of real numbers and substituting something new. In actual fact very little explicit use is made of this property in the text.

The representation of real numbers as points on a line enables the student to think geometrically about numbers. This ability is most useful since most of our basic concepts are arrived at geometrically and because the clearest way of thinking of functions is by means of their graphs.

The separation axiom, not being too well known, requires a little explanation. The statement is:

111

If A and B are non-empty sets of numbers having the property that every member of A is less than or equal to every member of B then there is a number s which separates A and B (i.e., $x \in A$ and $y \in B \longrightarrow x \leq s \leq y$).

This axiom bears a superficial resemblance to the Dedekind cut completeness axiom, but there are two basic differences, which are:

(1) A and B need not be intervals.

(2) A and B need not contain points arbitrarily close together.

The advantages of this axiom for our purposes are:

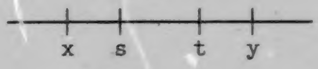
(1) These fewer hypotheses on the sets A and B which need be verified before the axiom can be invoked. This should make it much easier for the student to understand what the axiom says.

(2) It would seem that stated in the form the axiom has a high degree of intuitive appeal.

(3) This is the most useful form of the completeness axiom for the approach to integration adopted in Chapter 6. It is better than the least upper bound since we are not introducing upper and lower integrals in our approach. It is better than the Dedekind cut since, as soon as we know that every lower sum is less than or equal to every upper sum, we can invoke the separation axiom to discover that there must be a number separating the upper and lower sums. Showing uniqueness is a separate problem, and thus the difficulties are isolated.

It is easily proved that, if in addition to the hypothesis that "every member of A is less than every member of B" we also have "for every $\epsilon > 0$ there can be found numbers x in A and y in B so that $y - x < \epsilon$", then the number s separating A and B is unique.

The proof can be given as follows: Let $\epsilon > 0$ and let $x \in A$ and $y \in B$ with $y - x < \epsilon$. Since separation points s and t must both lie between x and y



it is clear that $|s - t| < \epsilon$. Thus $|s - t| < \epsilon$ for every positive number, so that $|s - t|$ must be equal to zero.

Further discussion of the use of this axiom will be found in the Explanatory Remarks for Chapter 6.

Some people with whom we have talked would prefer to eliminate all mention of completeness from this chapter and to broach the subject only in Chapter 6 when we make strong use of the idea. I feel that there are two major objections to this plan. First, we discuss real numbers here, but only mention properties which are also true of the rational numbers. The student will ask--if the real numbers are the same as the rational numbers why call them by a different name, and if they are not the same as the rational numbers, why isn't it explained how they are different? Second, the discussion of completeness in the chapter on the integral would constitute an ugly interruption in the flow of thought.

2.2 Inequalities and Absolute Value

As some of the attached materials show, this topic can very nicely be combined with the order properties of the real numbers. One basic idea is the triangle inequality. Some other ideas which must be stressed follow:

(1) $|a - b|$ is the distance between a and b on the number line.

(2) The set of numbers x satisfying $|x - a| < b$ is an interval with center at a and "radius" b.

(3) The set of numbers x satisfying $0 < |x - a| < b$ is the same interval as above with the center a deleted.

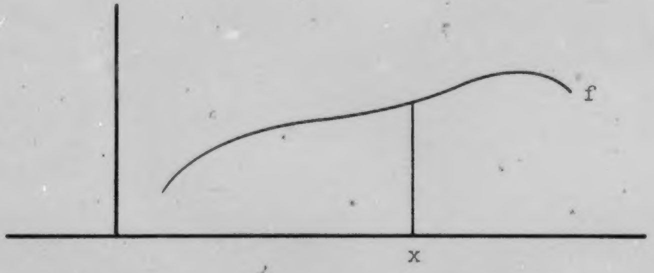
In order to get the idea in (3) across it may be advisable to discuss the set of numbers x satisfying

$$c < |x - a| < b,$$

and to show that $0 \leq |x - a| < b$ means exactly the same thing as $|x - a| < b$.

2.3 Functions

The first problem here is the concept of function--principally the idea of unique determination. There is some associated written material on this subject. Further expansion of this material should emphasize the distinction between f and $f(x)$ and illustrate this distinction geometrically. That is, in the illustration the function f is represented by the entire graph, while for each num-



ber x (in the domain of f) $f(x)$ is the ordinate of the unique point on the graph whose abscissa is x. Further emphasis should be made of the fact that a function is determined when the domain is given and a rule is given for finding $f(x)$ for each number x in the domain. This rule is very frequently a formula. Thus, for example,

$$g(x) = x^2 \qquad x > 2$$

defines a function g with domain consisting of the set of numbers greater than 2. If no explicit mention of the domain is made, then it is assumed to consist of all those numbers for which the formula makes sense. Thus if we write

$$s(x) = \sqrt{1 - x^2} ,$$

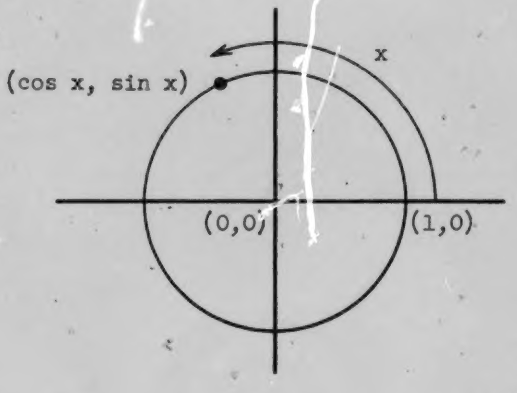
it is assumed that the domain of s is the interval $[-1,1]$.

2.4 Some Types of Functions

Increasing, decreasing, monotonely increasing, monotonely decreasing functions. Composite functions (perhaps also $f + g$ and $f \cdot g$). Inverse functions. Here it should be brought out that we restrict the discussion of inverses to those functions which are strictly increasing or strictly decreasing (since to all intents and purposes we propose to deal only with continuous functions).

2.5 Special Functions

Here just a little review is needed. It is assumed that this subject is part of the student's background. Slope should be emphasized in this review. Some polynomial functions could be graphed. Trigonometric functions must also be reveiwed with emphasis on the following definition.



When you measure a distance x around the unit circle counterclockwise from the point $(1,0)$, the coordinates of the terminal point are respectively $\cos x$ and $\sin x$. A few properties of trigonometric functions should be reviewed and the graphs of the functions plotted. This might be a good point to derive the inequalities.

(1) $\sin x < x < \tan x$ for $0 < x < \frac{\pi}{2}$, and the co-segment inequalities.

(2) $1 > \frac{\sin x}{x} > \cos x$ for $0 < x < \frac{\pi}{2}$.

There is a little motivation for doing this since the chapter deals with both trigonometric functions and inequalities. One weakness in this argument is that this particular inequality is not well motivated. Perhaps this objection can be overcome by presenting only the inequalities (1) and leaving (2) until we are ready to treat the derivative of the sine functions.

Explanatory Remarks on Chapter 2--
The Derivative

This chapter might better be titled: "Introduction to Derivatives and Limits." It should be explained that this material is very sketchy. It must have more detailed explanations and more examples. Furthermore, exercises must be added. The text must be expanded to adjust reading pace. The purpose of these written materials is to bring out a few main points and to indicate the intended scope and order of this short chapter.

The first idea is to introduce the concept and the definition of the derivative, starting from some geometrical or physical example. I feel that the first example should be geometrical so that the student can have some picture of what is going on to follow through these early stages of the development. I further feel that the example should be that of the tangent line (in Cartesian coordinates) because the student has the tendency to attach himself to the first interpretation of any concept which may be presented to him. The interpretation of the derivative as the slope of the tangent line is one which we always wish to have at our disposal in explaining problems in which the derivative occurs, no matter what the origin of the problem may be. This, of course, cannot be said of such an interpretation as velocity.

Many teachers seem to feel that the problem of the tangent line is not particularly interesting to students. If this is the case, then there should be some pre-motivation to make the student feel that tangent lines are important or interesting.

After introducing the problem of the tangent line to the graph of $f(x) = x^2$ at $(1,2)$, the student is shown that the slope of this line is given by

$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = \lim_{x \to 1} \frac{x - 1}{x - 1}(x + 1) \quad \text{for} \quad x \neq 1.$$

I should like to call attention to the use of $\frac{x - 1}{x - 1}(x + 1)$ instead of just $x + 1$. I feel that it might be quite a useful pedagogical device. After the first examples the student sees of differentiation in most texts, he probably would explain what is meant by a limit as follows: "First you divide out $x - 1$ numerator and denominator, and then you let x equal 1." The student must think: this process is rather mysterious but it always works. And so it does in his earliest examples. We would like to emphasize the idea of thinking of assigning values to x which are close to 1 rather than thinking of letting x equal 1 after cancelling. The basic difference in the two expressions:

$$\text{"}\frac{x^2 - 1}{x - 1} = x + 1 \qquad \text{for } x \neq 1\text{"}$$

and

$$\text{"}\frac{x^2 - 1}{x - 1} = \frac{x - 1}{x - 1}(x + 1) \qquad \text{for } x \neq 1\text{"}$$

so that in the second form the expressions on both sides of the equation are meaningless at 1, whereas in the first equation one side is meaningless and the other is not. The student may feel, in fact I am sure he often does, that our idea of limit consists of starting out with an expression meaningless for $x = 1$, dividing out this $x - 1$ top and bottom, and then conveniently forgetting that the original expression had no meaning for x equal to 1, and so substituting in the number 1 for x to get the answer. Our device for avoiding this process will not be completely successful since the student will soon notice that he can get the answer by substituting into the second factor in all those problems. However, if we can, even for a while, keep the student thinking in terms of values of x near to 1 rather than of letting x equal 1, I think we may be far ahead in eventually getting across to him the idea of what we mean by a limit.

There follow the definition of the derivative and then a number of examples elaborating the technique used in the first example. When we come to such an example as finding $f'(x)$ where $f(x) = \sqrt[3]{x}$, it should be stressed in the text that the student is not expected to have had the ingenuity to have thought of the rationalizing trick involved. The purpose of such an example is to plant the feeling of the need for general formulas and theory for derivatives to replace the method of returning to the definition and finding a suitable trick every time some new problem comes up.

The last example in the chapter is that of finding the derivative of the sine function which leads to the limit

$$\lim_{x \to 0} \frac{\sin x}{x}$$

which is evaluated by the usual area technique. The importance of this example lies in the impossibility of cancelling out the x and then substituting 0 for x. It is now my feeling that it would have been better to have gone as far as deriving the inequality $\sin x < x < \tan x$ in Chapter 2. Several teachers have remarked that they feel that this problem is too difficult for inclusion at this point. I am at a loss to understand this remark, especially as the same teachers have opined that the students will be able to handle ϵ's and δ's, which are scheduled to start on the page of the text following the one under discussion. I am loathe to part with this example at this point for a reason which I will explain in the next paragraph. If, however, it is indeed too difficult, then I propose the alternative of finding $f'(0)$ when $f(x) = \sin x$ instead of $f'(x)$. This would lead directly to the limit

$$\lim_{x\to 0} \frac{\sin x}{x}$$

instead of to the more complicated limit

$$\cos x \cdot \lim_{h\to 0} \frac{\sin h}{h} - \sin x \cdot \lim_{h\to 0} \frac{1 - \cos h}{h}.$$

Following the computation of $f'(0)$ we could take up the harder problem of $f'(x)$.

I will now explain why I consider this example to be of such vital importance pedagogically. If the student feels that he is able to evaluate all the limits he will ever encounter by means of this "cancelling out the h" device (perhaps applied after some algebraic trick), then he just won't hold still for the ϵ-δ treatment which is coming up in the next chapter. Only after he has seen that his technique is not universal is there any chance that he will give his attention to the formal theory. I do not feel that this purpose will be adequately served by raising the question of finding the derivative of $\sin x$ and not answering it at this point. I feel that in such a case the student would not necessarily believe that there is a limit; or he might feel that if there is one the best we could possibly do with it is to leave it expressed in the geometrical form as the slope of the tangent line. If it were decided that ϵ-δ should not be taken up in the text, then the argument for including this derivative at this juncture would lose most of its force.

According to the overall plan, this chapter has to end here. That is not to say that there should not be a number of interesting geometrical and physical applications of the ideas already developed. What we mean by saying that the chapter has to end here is that there is now no place to go with the theory of differentiation. We cannot, for example, now proceed with theorems on the derivative of the sum and product, etc. The reason is that we have decided to present the ϵ-δ treatment of limits.

Suppose we should at this point go on to the above-mentioned derivative theorems using only the intuitive concept of limit. That is, we offer arguments to make the necessary limit theorems seem plausible, and then use these limit theorems to prove the various derivative theorems. There is nothing wrong with this attack on the problem. I myself believe that it is, in fact, the best way to handle this problem at this level.

However, suppose we now follow such an exposition with an ϵ-δ treatment. Look what happens. The student is asked to try to master the most difficult mathematics he has ever been exposed to for the sole purpose of supplying proofs for theorems which he already believes! In fact, he has more faith in the theorems themselves than he has in the material that is being introduced in order to prove them. We seem to be saying, "Perhaps you had some doubts about these theorems. Perhaps the following explanation will convince you." In the usual case in which the student cannot understand the explanation, we will only have succeeded in casting doubt on the theorems that the student was ready to accept. He certainly will not see why the ϵ-δ theory was necessary.

A change in order considerably changes the situation. We start out from the indicated point of departure in the theory of the derivative and explain that in order to understand what is going on in this matter of limits we need a more precise definition of limit. We make an initial apology for the difficulty of the concept. We come up with the definition and then use it to prove the necessary limit theorems. Then comes the final apology. This consists of saying that if he can't follow the intricacies of the ϵ-δ treatment, then all he has to do is to accept the limit theorems on faith as there will be no further use of ϵ and δ. This doesn't seem too much to demand as these theorems seem quite reasonable anyhow.

With this method of presentation it is hoped that the student will see that all this work on ϵ and δ serves some useful purpose and is not just a pointless aside.

Explanatory Remarks on Chapter 4-- Limits

As mentioned twice before, I am not convinced of the wisdom of including an ϵ-δ treatment of limits in such a text. I would withdraw my objections if I felt that the students would understand any considerable portion of such an exposition. It is generally agreed that very few students develop any real understanding of ϵ-δ technique the first time around. Many experienced teachers feel that students begin to have an appreciation for the ideas involved only after the third exposure. For this reason I would also withdraw my objections to includion of ϵ and δ if the student is not left in a state of confusion and hostility. Perhaps this goal can be attained, after all.

Before attempting to make any improvement in the method of explaining ϵ and δ, one must first attempt to discover where the difficulties lie. One does not have to look very far to find them. Some of these problems are enumerated below:

(1) The first problem is best appreciated by looking at the definition of $\lim\limits_{x\to a} f(x) = L$ in formal notation.

$$(\lim_{x\to a} f(x) = L) \longleftrightarrow \underset{\epsilon>0}{\forall}\ \underset{\delta>0}{\exists}\ \underset{x}{\forall}\ (0 < |x - a| < \delta \rightarrow |f(x) - L| < \epsilon).$$

Here one sees that the definition contains three quantifiers followed by an implication. One can alter the wording to be more acceptable, but the three quantifiers are still there. They won't go away. The importance of the order of quantification is very difficult to get across.

(2) A second difficulty is due to the fact that we are not defining the expression $\lim\limits_{x\to a} f(x)$, but instead we are defining the statement $\lim\limits_{x\to a} f(x) = L$. Furthermore, this statement involves the use of equality before we know that the limit is unique. In short, when we make the definition, we are not sure that this definition is consistent because of this use of equality.

(3) The definition does not (although the subsequent theorems do) offer the student a method of finding the value of $\lim\limits_{x\to a} f(x)$ in any particu- instance but instead affords him a method of checking or verifying a conjectured answer found by some other method. Students seem to find

such a method of thinking most foreign to them. I believe that this is one of the reasons that students have so much trouble with mathematical induction where the same difficulty is encountered. The student always wants to know where that answer came from. If you can show him how to get the answer then he wants to know why further proof is necessary.

(4) This difficulty is somewhat like number (3). Many treatments of limits start out with the intuitive concept, then leap to the formal definition, finally demonstrate that the formal definition embodies the ideas of the intuitive concept. The student wants to see things done the other way around. He wants to see how to start from the intuitive idea and arrive at the formal definition. "How did you get that definition?" the student asks.

(5) The technique involved in using ϵ and δ in the simplest examples involves ingenious tricks, especially that of the two (or several) restrictions on δ.

(6) Further difficulty with the technique is encountered in that we do not solve for δ in terms of ϵ or try to find the best δ (which would entail severe difficulties even in the simplest examples). Instead we look only for a δ which will work. There is no unique answer. There is also a problem in understanding that if a δ will work for a given value of ϵ, then so will any smaller value of δ, and that if a δ will work for a given value of ϵ, then this δ will also work for any larger value of ϵ.

Now we come to the point of discussing the attached student materials. In the first place it should be remembered that it was decided to motivate limits through derivatives. Derivatives motivate the limits of functions defined on intervals and not limits of sequences. The only motivation available at this point for studying limits of sequences lies in the use of the same word, "limit," to describe these two phenomena. It is no doubt true that starting the subject of limits with limits of sequences has some pedagogical merit, especially as the idea of the limit of a geometrical series lies so close to the student's experience. On the other side of the ledger we have the confusion involved in having two kinds of limits and the additional class time involved in teaching two kinds of limits. Furthermore, as this is a one-year text, infinite series and sequences will not be included, so that no further use is made of the limit of a sequence in this course. If limits of sequences must be included, then it would be preferable to abandon the idea of using the derivative to motivate the study of limits and revert to the order of having a chapter on limits precede the introduction of the concept of the derivative.

It will be seen that in the attached materials a great effort is made to handle problem (4) above. In the course of working on this problem the author came to the realization that the "natural" transition from limits intuitive to limits rigorous consists of a series of questions and answers through which we are gradually forced to come to grips with the problem and refine our statements. Accordingly this writer tried his hand at bringing out this transition by means of a conversation between two students. Halfway through, the writer realized that this material was too drawn out for the text but that it might make--after expansion and polishing--a suitable scenario for a supplementary film on limits. On the chance that some of the ideas therein may be useful to writers, this material is included under the title of "Limits Supplement."

Returning to the list of problems in the presentation of limits, I see no better way to attack problem (1), that of the three quantifiers, than by making strong use of and frequent return to the geometrical interpretation of the definition of limit. Here one can actually see what is meant by "for every $\epsilon > 0$, there is a $\delta > 0$, etc." Much more can be done with this idea than is done in the material offered herewith. Part of the difficulty encountered in problem (2) above can be avoided by using the notation $f(x) \to L$ as $x \to a$ until the uniqueness of the limit has been proved, thus avoiding the possibility of inconsistency being introduced by incorrect use of equality. An effort has been made to partially meet problem (5) by considering as the first example of limit technique the limit $\lim_{x \to 1} \frac{1}{1 + x^2}$. In this example the "other factor" is universally bounded, so that there is no need for two restrictions on δ. Many people have remarked, and I am forced to admit, that there are so many other complications involved in this particular limit that it can hardly be regarded as simple. I am quite willing to accept the substitute $\lim_{x \to 0} \frac{1}{1 + x^2}$, which again has the "other factor" of being universally bounded. The proof that $\lim_{x \to 0} \frac{1}{1 + x^2} = 1$ is given here for comparison.

$$\text{If } 0 < |x| < \delta, \text{ then } \left| \frac{1}{1 + x^2} - 1 \right| = \frac{x^2}{1 + x^2} < \delta^2 \frac{1}{1 + x^2} \leq \delta^2 .$$

Therefore, if $\epsilon > 0$ and we choose $\delta = \sqrt{\epsilon}$, we have

$$\text{if } 0 < |x| < \delta, \text{ then } \left| \frac{1}{1 + x^2} - 1 \right| < \delta^2 = \epsilon .$$

Nothing has been done in the accompanying student materials about problem (3), and what little has been done about problem (6) needs no further discussion.

For lack of time, nothing more was written of this chapter. The following are the theorems and definitions which need to appear here for the purposes of the remainder of the book. They are given more or less in the order in which they should appear. I offer no suggestions as to which should be proved and which (if any) should merely be stated without proof.

(1) Limit of constant;

(2) Limit of sum;

(3) Limit of product;

(4) Limit of quotient;

(5) Squeeze theorem (if $f(x) \leq g(x) \leq h(x)$ and $\lim_{x\to a} f(x) = \lim_{x\to a} h(x)$, then $\lim_{x\to a} g(x)$ exists and $\lim_{x\to a} g(x) = \lim_{x\to a} f(x)$;

(6) Definition of continuity at a point;

[Please include here the fact that the definition of limit with "<0" deleted in "$0 < |x - a| < \delta$" coincides with the definition of continuity.]

(7) Composition theorem;

[Here there are two forms:

(i) if $\lim_{x\to a} g(x) = b$ and f is continuous at b, then $\lim_{x\to a} f\big(g(x)\big)$ exists and is equal to $\lim_{y\to b} f(y)$.

(ii) if $\lim_{x\to a} g(x) = b$ and $\lim_{y\to b} f(y) = b$ and g does not assume the value b in some deleted nbhd of a, then $\lim_{x\to a} f\big(g(x)\big)$ exists and is equal to $\lim_{y\to b} f(y)$.

As a matter of fact, if $\lim_{x\to a} g(x) = b$ and $\lim_{y\to b} f(y)$ exist, then $\lim_{x\to a} f\big(g(x)\big)$ exists and is equal to $\lim_{y\to b} f(y)$ iff either (α) f is continuous at b or (β) g does not assume the value b in some deleted neighborhood of a. I feel that only form (i) should be included in the student's text and that in that case the conclusion should be stated in the form given above as well as in the form $\lim_{x\to a} f\big(g(x)\big) = f\Big(\lim_{x\to a} g(x)\Big)$ where it appears as a sort of commutivity theorem. Warning: this form does not hold in case (ii). I do think that it might be well to include the additional remark surrounding this theorem in the T.C. in order to clarify the hypotheses in the case we actually use.]

It will be assumed when we come to the chain rule that this theorem has been presented here. The proof of this theorem is most simple, being a simple syllogism, to wit: Let $\epsilon > 0$. Choose $\eta > 0$ so that $|y - b| < \eta \rightarrow |f(y) - f(b)| < \epsilon$. Now $0 < |x - a| < \delta \rightarrow |g(x) - b| < \eta \rightarrow |f\big(g(x)\big) - f(b)| < \epsilon$. I would suggest that this be one of the theorems to be actually proved in the text for the reasons that: one, the proof is very simple; two, the result may not seem to be obvious to the student because of the apparent mystery of the necessity of the additional hypothesis of continuity on f.

(8) The following useful corollary of (5) and (7): if f is continuous at a and $h(x)$ is such that $h(x)$ lies between a and x, then $\lim_{x\to a} f\big(h(x)\big) = f(a)$.

(9) Definition of continuity on a set.

(10) Statements (but not proofs!) of the maximum theorem and intermediate-value theorem for a continuous function on a closed interval.

It might be well to point out the global nature of these last theorems as distinguished from the local nature of the other theorems. The proofs of these theorems from the completeness properties of the real numbers should probably be included in the T.C.

(11) Without much fuss, present unilateral limits and continuity and state theorems without proof.

The chapter should end with a statement to the effect that the ideas met in this chapter are admittedly difficult and mastered only on repeated exposure. However, all we ask of the student is that he believe the theorems enumerated above. Deriving these theorems is the only purpose to which ϵ and δ will be put in this book.

Explanatory Remarks on Chapter 5-- Theory and Applications of the Derivative

The only student materials submitted for this chapter are those on the chain rule. After a great deal more work than the problem merits, we came to the conclusion that the proof given here is the only one which is at the same time rigorous and comprehensible to the student. This proof does entail the prior explanation that the existence of $f'(a)$ implies the existence of a function μ with

$$\lim_{x \to a} \mu(x) = \mu(a) = 0$$

such that

$$f(x) - f(a) = f'(a)(x - a) + \mu(x)(x - a).$$

The continuity of the function μ at a is most important, as it is this which allows us to apply our composition theorem for limits. The presentation of this theorem, starting as it does with a realistic problem, is characteristic of the style of presentation we should like to see used throughout the book. To find a realistic example which could be presented at this point, where the derivative of the composite function has a clear physical meaning and where neither of the component function is linear, was a most difficult task. This is the best we were able to do.

One necessary remark: It was desired to hold down the number of applications to the bare minimum necessary to drive home the meaning and importance of the derivative. It is desired to proceed as rapidly as possible to the integral, fundamental theorem, and logarithmic and exponential functions in order to have these functions available before all the applications are used up.

The applications of curve plotting, velocity, and extrema should serve to drive home the geometrical meaning of the derivative, its physical application, and its power in problem solving.

The formal theory of the derivative included in this chapter consists of

(i) the elementary combination formulas: sum, product, quotient, chain rule, inverse function formula;

(ii) formulas for differentiation of polynomials, roots, trig and inverse trig functions;

(iii) Rolle's theorem and law of the mean.

The law of the mean is needed in this chapter for two reasons. First, it is needed in curve plotting for the conclusion that a function whose derivative is positive in an interval is increasing in that interval. Second, it is needed in

Chapter 7 on the FTC where we need to know for one form of the FTC that a function is constant if its derivative is identically zero. Similarly, the inverse function formula is needed for two reasons. First, we need it in this chapter for derivatives of roots and of the inverse trig functions. Second, it is needed in Chapter 8 in the theory of the exponential function. I would suggest assuming the existence of the derivative of the inverse function after some heuristical geometrical justification.

A reminder is in order here. Recall that we have decided to suppress all notation involving variables until the introduction of the concept of the differential in a later chapter. Therefore, we have only the $f(x)$ notation for functions and the $f'(x)$ and $D_x f(x)$ notation for derivatives. At a later point we will introduce such notations as $\frac{dy}{dx}$, y', $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$. Also at this time we will see how the substitution theorem for integrals looks in terms of the notation of differentials.

Explanatory Remarks for Chapter 6--
Area and Integral

Two drafts of the student materials for this chapter are submitted. An original draft of most of this chapter was written which was subsequently rewritten in more or less final form and somewhat revised. This revision was not completed and, as the revision uses different notation than the original, it is difficult to switch back to the original from the revision after the latter breaks off. Both drafts have therefore been included.

There are a number of novel ideas involved in the approach to the integral adopted in this chapter.

One of these features is that the integral is approached through upper and lower sums (instead of Riemann sums), but we go directly to the integral and do not bring in upper and lower integrals. The reason for wanting to do things in this way is manifold. First, the idea of the integral is introduced and motivated by the problem of the area under a curve, and it is highly intuitive to get at area by approximating from above and below. Second, we feel with many of the people we have consulted that upper and lower sums are always the clearest way to explain the integral. Third, the upper and lower sum treatment provides an easy way of giving bounds on the error of approximation, bounds which are not available with a Riemann-sum approach unless you do something equivalent to introducing upper and lower sums.

The upper- and lower-sum attack is pursued without reference to supremum and infimum. The idea of the least upper bound is not one which students have much success in working with the first time around in calculus. We have therefore stuck entirely to maximum and minimum instead of infimum and supremum. We achieve this desired state of affairs honestly by treating only continuous functions in our discussion of integration, so that our integrals always attain a maximum and a minimum on a closed interval. (This fact is stated but not proved in the chapter on limits.) In by-passing the upper and lower integrals we again avoid the necessity of considering the supremum and infimum.

The secret that enables us to skirt the upper and lower integrals lies in the choice of the form at the completeness axiom adopted in Chapter 2. This seldom-seen form of the completeness axiom is but extremely intuitive and directly applicable. It can be expressed in several ways:

(a) If A and B are non-empty sets of numbers such that every member of A is less than or equal to every member of B, then there is a number s such that every member of A is less than or equal to s and every member of B is greater than or equal to s. [s may be a member of none, one, or both of the sets A and B.]

(b) if A and B are non-empty sets of numbers such that every member of A is less than or equal to every member of B, then there is a number which separates A and B. [s separates A and B is $x \leq s \leq y$ for all x in A and all y in B.]

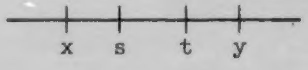
(c) $\{(x \in A \wedge y \in B) \rightarrow x \leq y\} \rightarrow \underset{s}{\exists} \{(x \in A \wedge y \in B) \rightarrow x \leq S \leq y\}$.

(As seen in c) there is no actual need to specify that A and B are non-empty (or even that they are sets of numbers), but the confusion which would arise and the odd meaning of the definition in such cases should be avoided by specifying that A and B are non-empty sets of numbers.

Once this axiom has been adopted it is very easy to prove that, if in addition to this property A and B have the property that these are members of A and B arbitrarily close together, then the separating number is unique. There is the way it goes.

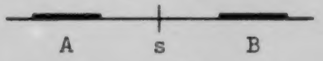
Theorem. Let A and B be non-empty sets of numbers for which every member of A is less than every member of B. Suppose further that for every $\epsilon > 0$ there exist numbers $x \in A$ and $y \in B$ so that $y - x < \epsilon$. Then the number separating A and B is unique.

Proof. Let s and t separate A and B. Let $\epsilon < 0$ and let $x \in A$ and $y \in B$ with $y - x < \epsilon$. From the meaning of separation we have $x \leq s \leq y$ and $x \leq t \leq y$ or, in other words, both s and t lie between x and y. There-



fore, s and t must differ by less than ϵ. Now we have $|s - t| < \epsilon$ for all positive ϵ. It therefore follows that $s = t$.

The advantage of this axiom over other forms, apart from its application to the definite integral lies in the fact that the difficulties have been separated, the hypotheses on A and B and the resulting conclusion are extremely simple and highly intuitive. The student will first picture A and B as quite far apart and will certainly agree that there is a point separating them. When A



and B are close together, he still will not experience any difficulty, not even when they are "touching." He will soon find that he has agreed to more than he realizes, of course, but this should not be regarded as a drawback. This experience is naturally typical of what happens in the axiomatic development of any mathematical field; it is what happens when we discover a remarkable theorem, or for that matter any non-intuitive theorem. What one wants to do when axiomatizing any mathematical topic for which the reader already has a strong intuition is to adopt axioms which the reader will certainly agree are formulations of his intuitive ideas. Having accepted these axioms, the reader has to stick with them when the going gets rough (i.e., when unexpected or non-intuitive theorems start showing up). If the student had any reservations about the axioms at the beginning, then when these unexpected results appear he may react with, "You made those initial assumptions, I didn't."

The problem then with the Dedekind cut or least upper bound approach is that the statements of those axioms immediately focus attention on the difficulties. The real problem is that the existence and uniqueness of the cut point or the supremum are both involved in the axiom even if the uniqueness is not stated. That is, in both approaches the number guaranteed by the axiom actually is unique. This is not the case with our separation axiom; the student doesn't get involved with the uniqueness of the separation number in considering the axiom; sometimes this number is unique and sometimes it isn't. Curiously enough in our approach to integration this non-uniqueness is very useful in that we are able to invoke the axiom at a time and in a way which would otherwise be impossible. We will now go on to see how this axiom is used in the development of the integral.

The Dedekind-cut axiom is cumbersome to apply to the problem of the integral primarily because of the requirement that to get a cut point between A and B from the axiom it is necessary to show that $A \cup B$ is the entire set of real numbers. The supremum axiom is made to order to apply in developing the integral when the upper and lower integrals are to be developed first. (In advanced courses, where it is desired to develop the concept of the Darboux integral for functions which are only known to be continuous almost everywhere, the supremum and infimum are still indispensible.)

In the course at hand the introduction to the idea of integral is conceived of as a gradual transition between two conceptual states--from the state of basing the idea of the integral on the intuitive properties of area to the state of basing the integral on properties of the number system. In this transition one encounters the ticklish problem of the true nature of intuition concerning area. We feel that the intuitive connection between area and number is very tenuous except in the case of rectangles and triangles. The student certainly feels that regions have areas, but it seems not obvious to him that there is actually a number greater than all the lower sums and less than all the upper sums. He rather seems to think of area as a partial ordering of regions by inclusion and with an additional idea of equivalence, i.e., that two sets "have the same area" when they can be decomposed into pairwise congruent subsets. [The student probably would be surprised to realize that this is the way he is thinking of things.] Support for this hypothesis can be obtained by studying the attitude that the Greeks had toward area. In consequence of this problem of intuition one must be careful to make it clear that what we mean by area is a real number.

In the course of the above described "transition" it is shown that every lower sum is less than or equal to every upper sum. (The possibility of equality occurs only with constant functions--not, for example, with step functions.) At this point we have two sets of numbers, $\mathcal{L}$ and $\mathcal{U}$, the sets, respectively, of lower and upper sums for the function over the interval to which the separation theorem applies. We invoke the axiom and thus find that there is a number (and perhaps many) separating $\mathcal{L}$ and $\mathcal{U}$. The question now arises as to whether the properties of area which we chose are adequate to uniquely determine the area under the graph of a continuous function. This brings us to ask whether we can show that the number separating $\mathcal{L}$ and $\mathcal{U}$ is unique. Done in this way we have not had to establish the condition for uniqueness before invoking the completeness axiom. We feel that separating these two problems makes for a clearer presentation. It is also well to note here that if we had used the supremum form of completeness, then this is the point at which the extraneous concepts of upper and lower integrals would appear. We must in all fairness admit that it would not be necessary to give them these names. We could instead speak of the least upper bound of the set of lower sums and the greatest lower bound of the set of upper sums.

In considering the problem of the uniqueness of the number separating and , we look first at continuous monotone functions. Here a part of our strategy is revealed. In an earlier section in connection with the problem of finding a priori bounds on the error in approximating the area by upper and lower sums, we had found that for continuous monotone functions

$$\overline{S}_{\Delta} - \underline{S}_{\Delta} \leq |f(b) - f(a)| \; ||\Delta||.$$

Consequently it requires merely a remark to see that by making $||\Delta||$ sufficiently small we may make $\overline{S}_{\Delta} - \underline{S}_{\Delta}$ as small as we wish. Then the uniqueness theorem for the separating number which is proved in these remarks can now be involved. A similar result is true for piecewise monotone continuous functions.

It is next stated that this uniqueness theorem in the case of continuous functions involves certain technicalities which we do not wish to go into. The theorem is then stated but not proved for continuous functions. It is remarked at this point in the text that the only functions which we will encounter in this text are functions which are piecewise monotone and for these functions the theorem has actually been proved. It is suggested in the text that the only reason for stating the result for continuous functions at this time, when we cannot prove it and will not use it, is the reason of keeping in step with the rest of the world. It was thought best not to reveal to the student the real reason that we have to state this theorem for continuous functions.

This reason is that the class of piecewise monotone continuous functions is not closed under addition. The smallest class of functions which contains the piecewise monotone continuous functions is the class of continuous functions of bounded variation. Every continuous function of bounded variation can, of course, be expressed as the sum of two monotone functions. This creates severe difficulties when we come to the theorem

$$\int_a^b (f + g)(x)dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$

What hypotheses can be used to precede this statement? There are two possible choices.

(i) If f and g are continuous on $[a,b]$, then;

(ii) If f, g, and $f + g$ are continuous and piecewise monotone on $[a,b]$, then.

The second choice has the disadvantage that it requires a condition on $f + g$ to be verified before the theorem can be used. The student is actually working within a still smaller class of functions which is closed under addition--the class of analytic functions.

This class cannot be described to the student at this time. The simplest way out of these difficulties would seem to be to make the true but unproved statement that the integral of a continuous function actually exists.

At the time at which these pages in the student materials were written, it had not been decided at which point the concept of the integral for functions which may assume negative values should be introduced. We feel that we now know the answer to this question. The reason for contemplating a delay in introducing this idea in the first place was in order to have at our disposal the interpretation of area with which to follow the proofs of the various theorems on integration. The proofs would be analytical, but they can be followed step by step with the area interpretation.

We now realize that it is possible to have it both ways! We will outline our proposed method of procedure.

As soon as the definition of the definite integral has been arrived at, we make the remark that this concept is also extremely useful for functions which assume negative values; although it is not evident now it soon will be. Therefore we define for any function continuous on $[a,b]$,

$$\int_a^b f(x)dx = \text{the unique number}$$

separating the sets $\mathcal{L}$ and $\mathcal{U}$ of lower and upper sums. This would be an excellent time to interject the remark that our concept of integral no longer depends in any way on the intuitive idea of area, in spite of having been inspired by this intuitive idea. The integral now rests entirely on properties of the real number system. To this end it is only necessary to show that neither the existence of a number c separating $\mathcal{L}$ and $\mathcal{U}$ nor its uniqueness depends on our intuition concerning area. This does not mean, however, that we will abandon area as an interpretation of the integral. [Here it might be pointed out in the T.C. but not in the student's text that the establishing of the existence and uniqueness of the integral as done here is, though less intuitive, much simpler than showing the consistency of our initial assumptions regarding area.]

Next we would consider a function f continuous on $[a,b]$ and let A be such a number that $f(x) + A \geq 0$ for all x in $[a,b]$. Let $g(x) = f(x) + A$ for x in $[a,b]$. Let $\Delta = \{x_0, x_1, \ldots, x_n\}$ be an arbitrary partition of $[a,b]$ and let M_1 and M_1' be, respectively, the maximum values of $f(x)$ and $g(x)$ in the k^{th} sub-interval. Now $M_k' = M_k + A$, $k = 1, 2, \ldots, n$. Letting $\overline{S}_\Delta = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$ and $\overline{S}_\Delta' = \sum_{k=1}^{n} M_k'(x_k - x_{k-1})$, we have

$$\begin{aligned}\overline{S}_\Delta' &= \sum_{k=1}^{n} M_k'(x_k - x_{k-1}) = \sum_{k=1}^{n} (M_k + A)(x_k - x_{k-1}) \\ &= \sum_{k=1}^{n} M_k(x_k - x_{k-1}) + \sum_{k=1}^{n} A(x_k - x_{k-1}) \\ &= \overline{S}_\Delta + A(b - a).\end{aligned}$$

Similarly, $\underline{S}_\Delta' = \underline{S}_\Delta + A(b - a)$. Thus for each partition Δ the upper and lower sums for g are exactly $A(b - a)$ more than the upper and lower sums for f. Therefore the unique number separating the upper and lower sums for g is just $A(b - a)$ more than the corresponding number for f. That is,

$$\int_a^b g(x)dx = \int_a^b f(x)dx + A(b - a).$$

This will enable us to prove the subsequent theorems first for non-negative functions and then obtain the general results as very simple corollaries.

In this way, then, the student can follow the essential parts of the proofs of all theorems with his area interpretation. See, for example, the proposed treatment of the FTC exhibited in the Explanatory Remarks for Chapter 7. I feel that there is a very important pedgogical point involved here. In many other expositions the student is in effect told that, in the case that the function f is positive, we have a geometrical interpretation of the theorem and its proof, while in the case that f is not positive, this interpretation breaks down and the student will have to rely on the analytical statements. With our approach the student can see why it is sufficient to prove these theorems for the positive case.

It will be observed on reading the student text that a very unusual presentation has been given of the summation formulas encountered in the early stages of the theory of integration. It consists in using "telescoping sums" throughout instead of the usual method of mathematical induction. We have a number of reasons for preferring our treatment. First, and most important, we do not wish to digress here to take up the topic of induction. Second, every summation formula presented at this level by means of induction in essence uses the technique presented here. Third, the student will not need induction to see that $\sum_{k=1}^{n} u_k - u_{k-1} = u_n - u_0$. Fourth, although the method is admittedly rather tricky, it at least affords the student a method of finding the answer (if only he can find the right trick) rather than just a method of checking the answer once it is given. Fifth, once the trick has been supplied, it is as easy (if not easier) to verify the summation formula by our method as by induction. [This also applies if the summation formula is given instead of the trick.] Sixth, this fresh approach will most likely be interesting and stimulating to the teacher.

A remark is in order on the inclusion of the result $\int_0^a \cos x \, dx = \sin a$ in this chapter. There are several reasons in support of its inclusion and one against it. I realize that the one reason against may outweigh all the reasons for. Perhaps, therefore, it should be included in a starred section. Among the reasons for is the desire to emphasize the definition of the integral as the limit of a sum over the interpretation as the difference in two values of the antiderivative. To this end we wish to do more than the usual amount of work using this definition and actually see that we can get some results of more than ordinary interest. In addition to emphasizing the definition of the integral, this example ought to serve the purpose of bringing out the power of the FTC when we come to it by showing how many difficulties it cuts through in this case. Finally, this example exhibits the importance of a trigonometric summation formula which the student will encounter over and over again in his future mathematical studies.

The final portion of this chapter should include the mean-value theorem for integrals and the theorem that, if $f(x) \leq g(x)$ for $a \leq x \leq b$, then

$$\int_a^b f(x)dx \leq \int_a^b (x)dx.$$

Explanatory Remarks for Chapter 7--
The Fundamental Theorem of Calculus

We regret that there has been no time to prepare sample student materials for this important chapter. We do, however, have rather definite ideas as to what should go in the chapter, and we will try to spell them out here.

It is of the greatest importance to make it clear that there are two forms of the Fundamental Theorem of Calculus, the first dealing with the derivative of the integral,

$$D_x \int_a^x f(t)dt = f(x), \quad f \text{ continuous},$$

and the second dealing with the integral of the derivative,

$$\int_a^x F'(t) = F(x) - F(a), \quad F' \text{ continuous}.$$

Together these theorems demonstrate the inverse nature of differentiation and integration. The student should learn to call either of these forms the Fundamental Theorem of Calculus. We hope that he will not favor one form over the other.

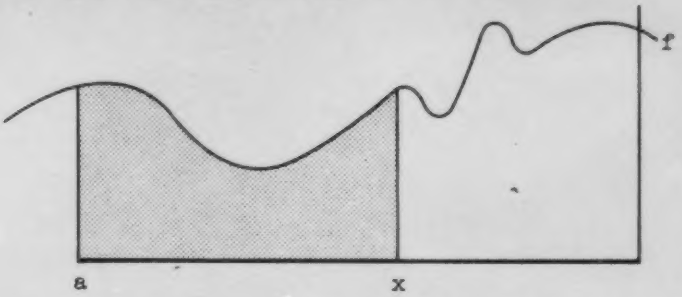
As for the method of presentation, I feel that it should be treated as follows. We start with the first form and restrict ourselves to non-negative functions f so as to have the interpretation of area available. Then we prove the theorem for functions which may assume negative values as a corollary of the theorem for non-negative functions. I further feel that the proof might be clearer if presented through unilateral limits. The proof will be analytical, but the student will have his geometrical area picture with which to follow the proof. There follows a sketch of how this might all be done. I do not know whether it is best to use the mean-value theorem for integrals in the proof or not. Probably not. The reason I say this is that, as will be recalled, the theory of the integral as developed in Chapter 6, insofar as it involved monotone or piecewise monotone continuous functions, make no use of any of the deeper properties of continuous functions. I would like to preserve this state of affairs in the proof of this theorem as well. That is, I should like to present the proof in such a way that, if the hypothesis "f is continuous" should be changed to "f is monotone (or piecewise montone) and continuous," then the proof would make no use of the deepar properties of continuous functions. The mean-value theorem for integrals makes use of the intermediate-value theorem for continuous functions. For the above reasons I will present two proofs here.

FTC (first form, first proof).

Suppose that f is continuous and non-negative on $[a,b]$. Let

$$F(x) = \int_a^x f(t)dt.$$

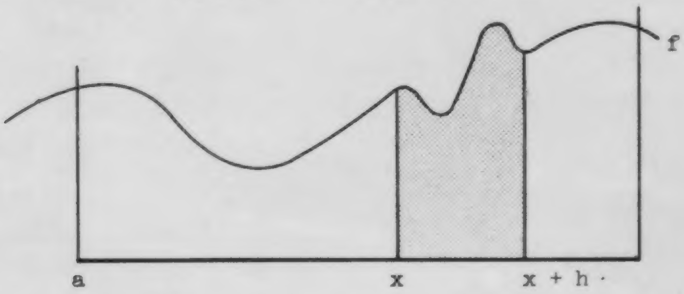
We can see that $F(x)$ then represent the area of the region under the graph of f between a and x as shown below.



We will try to compute the derivative $F'(x)$. That is, we will attempt to find the value of $\lim_{h\to 0} \frac{F(x + h) - F(x)}{h}$. Let x be some number between a and b, and let h be positive. We recall that

$$F(x + h) - F(x) = \int_a^{x+h} f(t)dt = \int_a^x f(t)dt = \int_x^{x+h} f(t)dt.$$

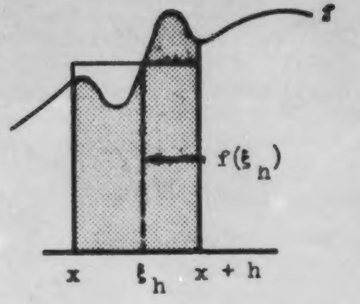
This last integral is represented graphically in the next figure.



(A) By the mean-value theorem for integrals

$$\int_x^{x+h} f(t)dt = f(\zeta_h)h$$

where $x < \zeta_h < x + h$. Thus the integral $\int_x^{x+h} f(t)dt$ is equal to the area of a rectangle with width h and height $f(\zeta_h)$.



So, we have seen that

$$F(x + h) - F(x) = f(\zeta_h)h$$

whence

$$\frac{F(x + h) - F(x)}{h} = f(\zeta_h)$$

so that

$$\lim_{h\to 0+} \frac{F(x + h) - F(x)}{h} = \lim_{h\to 0+} \cdot f(\zeta_h) = f(x).$$

This last comes from the composition theorem for limits (it is also easy to see from the graph). Since $\lim_{h\to 0+} \zeta_h = x$ and f is continuous at x, we have

$$\lim_{h\to 0+} f(\zeta_h) = f(x).$$

The same argument holds for negative h. We give the steps without explanation. For $h < 0$

$$\frac{F(x + h) - F(x)}{h} = \frac{\int_a^{x+h} f(t)dt - \int_a^x f(t)dt}{h} = \frac{-\int_{x+h}^x f(t)dt}{h} = \frac{-f(\zeta_h)\cdot(-h)}{h} = f(\zeta_h)$$

where $x + h < \zeta_h < x$. Thus

$$\lim_{h\to 0-} \frac{F(x + h) - F(x)}{h} = \lim_{h\to 0-} f(\zeta_h) = f(x).$$

Since both $\lim_{h\to 0+} \frac{F(x + h) - F(x)}{h} = f(x)$ and $\lim_{h\to 0-} \frac{F(x + h) - F(x)}{h} = f(x)$, we have

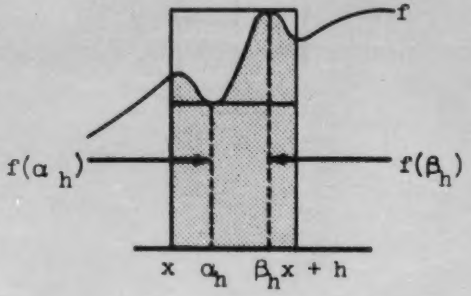
$$F'(x) = \lim_{h\to 0} \frac{F(x + h) - F(x)}{h} = f(x).$$

FTC (first form, second proof).

Start from (A) in the previous proof. Now

$$f(\alpha_h)\cdot h \le \int_x^{x+h} f(t)dt \le f(\beta_h)\cdot h$$

where $f(\alpha_h)$ and $f(\beta_h)$ are, respectively, the minimum and maximum values assumed by f in $[x, x + h]$. Graphically, then, the value of $\int_x^{x+h} f(t)dt$ lies between the areas of two rectangles, both with width h and with heights $f(\alpha_h)$ and $f(\beta_h)$.



We have seen, then, that

$$f(\alpha_h) \cdot h \leq F(x + h) - F(x) \leq f(\beta_h) \cdot h$$

so that

$$f(\alpha_h) \leq \frac{F(x + h) - F(x)}{h} \leq f(\beta_h).$$

Since α_h and β_h are both between x and $x + h$, it is clear by the squeeze (it is also clear from the graph) that $\lim_{h \to 0+} \alpha_h = x$ and $\lim_{h \to 0+} \beta_h = x$. Since f is continuous at x, we therefore find by the composition theorem for limits (again clear from the graph) that $\lim_{h \to 0+} f(\alpha_h) = f(x) = \lim_{h \to 0+} f(\beta_h)$. Again using the squeeze we see that

$$\lim_{h \to 0+} \frac{F(x + h) - F(x)}{h} \quad \text{exists}$$

and is equal to $f(x)$. The same result holds for negative values of h, etc.

Next we must consider the theorem for integrands which may assume negative values. Here is the way I should like to see used.

Suppose that f is continuous on $[a,b]$ but not necessarily positive. Find a constant A so that $f(x) + A$ is positive on $[a,b]$. Define $g(x) = f(x) + A$. If $F(x) = \int_a^x f(t)dt$, then

$$G(x) = \int_a^x g(t)dt = F(x) + A \cdot (x - a)$$

by a result of the last chapter. Now by the theorem just proved, $G'(x) = g(x)$, so that,

$$\begin{aligned} F'(x) &= D_x G(x) - D_x\big(A \cdot (x - a)\big) \\ &= g(x) - A = f(x). \end{aligned}$$

The condition of being non-negative may thus be removed from the FTC.

We next proceed to the proof of the second form of the FTC, which proof (alas) involves Rolle's theorem, which in turn depends on the maximum property of continuous functions.

After the inverse nature of differentiation has been brought out, the theorems

$$\int_a^b c\ f(x)dx = c\int_a^b f(x)dx$$

and

$$\int_a^b \big(f(x) + g(x)\big)dx = \int_a^b f(x)dx + \int_a^b g(x)dx$$

should be proved. It should be pointed out that since we are only interested in continuous integrands these theorems (especially the second) are much more easy to prove here with the FTC at our disposal than they would have been in the last chapter. Now they simply follow from the corresponding differentiation theorems. (I feel that the loss involved in not proving these theorems by techniques of integration is slight.)

The next order of business (perhaps it should even precede the last paragraph) is the definition of the indefinite integral or antiderivate. We point out that every differentiation theorem or formula can be expressed in terms of integrals. We find the integrals of x^n and of $\sin x$ and $\cos x$. Here we point out that the integral of $\cos x$ was computed in the previous chapter without the use of the FTC. We saw that this computation involved a great deal of effort and ingenuity. With the FTC at our disposal this result is entirely trivial.

Next we turn our attention to the form taken by the chain rule as an integral theorem. If

$$F(x) = s\big(g(x)\big)$$

then

$$F'(x) = s'\ g(x) \cdot g'(x).$$

Thus

$$\int_a^b F'(x)dx = F(b) - F(a) = s\big(g(b)\big) - s\big(g(a)\big)$$

but this last expression is readily recognized (again by the FTC) as being the value of $\int_{g(a)}^{g(b)} s'(u)du$. Thus we have (substituting f for s')

$$\int_{g(a)}^{g(b)} f(u)du = \int_a^b f\big(g(x)\big)\ g'(x)dx.$$

Further explanation of this theorem as a substitution theorem may be given. It should only be remembered that we do not yet have differentials at our disposal. If it is preferred to defer this theorem until later on, it can be done because it is not absolutely essential in the next chapter. Another method for obtaining the addition formula for the logarithm is included in that chapter.

I would now like to suggest a happy note on which this chapter might end. There are two valuable purposes which it will serve; it will give the student a striking impression of the importance of calculus, and it will show the usefulness of the technique of integrating inequalities, incidentally bringing out thereby an important distinction between definite and indefinite integrals. (It should be demonstrated in passing that we cannot differentiate inequalities.)

Almost all students must have wondered how these four- and five-place tables of trigonometric functions were compiled. Surely not by measuring the lengths of segments with a high degree of accuracy. Well, we now have at our disposal a method for quickly computing values of sines and cosines with any desired degree of accuracy. We start with the inequality

$$0 \leq \sin x \leq x \qquad \text{for} \quad 0 \leq x \leq \frac{\pi}{2}$$

already shown in Chapter 2 and used in Chapter 3. Integrating, we have

$$\int_0^t 0 \, dx \leq \int_0^t \sin x \, dx \leq \int_0^t x \, dx \qquad \text{for} \quad 0 \leq t \leq \frac{\pi}{2}$$

so that,

$$0 \leq 1 - \cos t \leq \frac{t^2}{2} \quad \text{or} \quad 1 - \frac{t^2}{2} \leq \cos t \leq 1 \qquad \text{for} \quad 0 \leq t \leq \frac{\pi}{2}.$$

Integrating again

$$\int_0^x \left(1 - \frac{t^2}{2}\right) dt \leq \int_0^x \cos t \, dt \leq \int_0^x 1 \, dt \qquad \text{for} \quad 0 \leq x \leq \frac{\pi}{2}$$

so that,

$$x - \frac{x^3}{3!} \leq \sin x \leq x \qquad \text{for} \quad 0 \leq x \leq \frac{\pi}{2}.$$

Another integration yields

$$\int_0^t \left(x - \frac{x^3}{3!}\right) dx \leq \int_0^t \sin x \, dx \leq \int_0^t x \, dx \qquad \text{for} \quad 0 \leq t \leq \frac{\pi}{2}$$

so that,

$$\frac{t^2}{2} - \frac{t^4}{4!} \leq 1 - \cos t \leq \frac{t^2}{2}$$

or

$$1 - \frac{t^2}{2} \leq \cos t \leq 1 - \frac{t^2}{2} + \frac{t^4}{4!} \qquad \text{for} \quad 0 \leq t \leq \frac{\pi}{2}.$$

Iterating this process (induction may be brought in if desired) we obtain, for example,

$$x - \frac{x^3}{3!} \leq \sin x \leq x - \frac{x^3}{3!} + \frac{x^5}{5!} \qquad \text{for} \quad 0 \leq x \leq \frac{\pi}{2}$$

and

$$1 - \frac{t^2}{2} + \frac{t^4}{4!} - \frac{t^6}{6!} \leq \cos t \leq 1 - \frac{t^2}{2} + \frac{t^4}{4!} \quad \text{for} \quad 0 \leq t \leq \frac{\pi}{2}.$$

We can use these results to obtain extremely accurate approximations of $\sin x$ and $\cos x$ for small values of x. It can be pointed out that the above inequalities give good estimates for $0 \leq x \leq \frac{\pi}{4}$ and that if we can tabulate these functions for such values of x, then all other values may be computed from simple trigonometric identities. This example also introduces the ideas of polynomial approximation.

Explanatory Remarks on Chapter 8--
The Logarithmic and Exponential Functions

There is little that is original in this chapter. The only reason for submitting student materials on this subject is that the writer wanted an opportunity to do something easy as a change from being constantly immersed in the most difficult problems of the calculus.

The main purpose of these remarks is to explain why the problem of the exponential and logarithmic functions ought to (we might almost say "has to") be attacked in this way. There are three basic reasons:

(i) the course is supposed to follow the SMSG Intermediate Mathematics course, which treats these functions in this way;

(ii) the alternative method of considering the exponential function first must either be very sketchy or involve a horrendous proliferation of the dullest possible theorems;

(iii) the suggested method of treating the logarithm as $\int_1^x \frac{1}{t}\,dt$ is very clean and brings all the machinery of integration (and even differentiation) to bear on a single problem, and thus ties together everything we have already done in a most elegant fashion.

It will be noted that the properties of these functions are not brought out in the theorem-proof-theorem-proof style but are instead presented informally in the text and finally collected at the ends of the sections.

Attention is called to the estimation of $L(x)$ for $1 \leq x \leq 2$ by the method of integrating inequalities.

115

Explanatory Remarks for Chapter 9 et. seq.

Virtually no thought has been given to the organization of the text from this point on. The Table of Contents offers one plan, but it is felt that this can be greatly improved upon when the problems of coordination and exposition are forced by future writing teams. I will conclude with a very small number of suggestions.

Almost everyone we have asked for suggestions has urged that in this age of computers numerical methods should be emphasized to a greater degree than is usual in calculus books. I would, therefore, urge the inclusion of Newton's method, the trapezoid rule, and Simpson's rule with bounds on error.

I feel that a good motivation for numerical integration would be furnished by arc length as, for all but the simplest (and a few special) functions, this leads to integrals which cannot be formally integrated.

Many consultants have suggested that formal integration techniques be somewhat de-emphasized. We pass this suggestion along.

Concerning arc length again, we suggest that convex (or piecewise convex) curves be concentrated upon, so that both lower and upper estimates for arc length can be obtained instead of just lower estimates.

It should be recalled that differentials have yet to be taken up and along with them various notations involving "variables" and a clarification of the substitution theorem for integration.

STUDENT TEXT

Chapter 1

INTRODUCTION
(first version)

An overly utilitarian view of the calculus is that it is merely a bag of tricks for obtaining useful solutions to abroad variety of scientific and technical problems. On the other hand, the calculus can be treated purely as an intellectual exercise, as a mathematical discipline in which theorems are deduced from carefully stated postulates and definitions, and then the primary question of interest is whether the logic is impeccable. In this text we try to maintain a more flexible point of view. We shall find the origins of the ideas of the calculus in practical problems; we shall attempt to express these ideas precisely so that we may reason about them logically; finally, we shall return to problems and apply the theorems resulting from our reasoning.

The two basic ideas of the elementary calculus are "derivative" and "integral." It is easy to appreciate these ideas intuitively and know why they are useful before formulating them precisely. Here we shall consider these ideas as they arise in the solution of specific problems.

1.1 The Concept of the Derivative

It is in the nature of the human enterprise to try to get the best of everything: a manufacturer seeks the smallest unit cost for his product and the highest possible price; a student tries to complete his homework assignment in the shortest possible time; a demagogue expounds the political philosophy which he believes will garner the greatest number of votes. It is seldom clear what must be done to get the best value; here we shall develop a systematic attack on a class of these problems.

While the class of "best value" problems treated by the methods of elementary calculus is quite broad, we are only making a beginning in an area which is still a lively field of investigation.

Consider the following problem which the writer faced recently in moving his household goods. The cost of shipping books by parcel post happened to be much lower than the cost of shipment by interstate van. The post office places restrictions on the size of packages: the length plus the girth must not exceed 72 inches. Since there were a great many books, to keep the effort of packing to a minimum the writer sought out the largest possible boxes complying with the post-office requirement. Assuming the ends of the box to be square*, what are the dimensions of the box of largest size?

To solve this problem we must know that the post office defines the girth of the box as the perimeter of the end piece. We let x denote the size of the square end and y the long dimension of the box; we require that

$$y + 4x = 72$$

and under this condition we attempt to maximize the volume of the box,

$$V = x^2y.$$

Setting $y = 72 - 4x$ in the expression for V we obtain

$$V = x^2(72 - 4x).$$

Getting away from the specific details, we see that what we have accomplished is to reduce the problem to the study of the properties of a function $f\colon x \to V$. Our problem is not so much to determine the largest value V_{max} in the range of the function, although that information may also be useful, but to find a value of a in the domain for which $f(a) = V_{max}$. (The domain here consists of those values of x for which the problem is meaningful; that is, the values between 0 and $\frac{72}{4}$.) In order to get some feeling for the problem, we may sketch the graph of f by plotting a few easily calculated points and drawing a smooth curve through them (Figure 1).

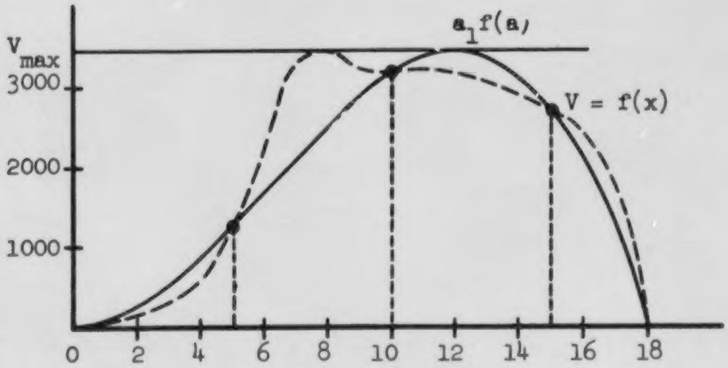


Figure 1. $x \to x^2(72 - 4x)$

In this way, we might locate a peak of the graph approximately, and we do get some precise information, such as $V_{max} \geq f(10) = 3200$. No matter how much information we get this way, we shall always be somewhat dissatisfied. In the first place, we have only exact information about the function at a number of calculated points, so that even if we happened to stumble upon the maximum we might not be aware of it.

---

*It is not hard to prove that the best box has square ends, but we shall postpone the argument for the sake of brevity here.

In the second place, the idea of drawing a smooth curve through the calculated points has its limitations. For example, in Figure 1, without further calculation we could not be sure that the continuously drawn curve more reasonably represented the function than the dashed one and, furthermore, we cannot eliminate this kind of ambiguity completely by calculating more points. One of our objectives is to devise systematic methods for resolving these difficulties.

Thinking of the problem in usual geometrical terms, we see that the condition for a maximum, $f(a) = V_{max}$, means that the graph of f cannot cross over the horizontal line through $[a,f(a)]$. The direction of the graph at $[a,f(a)]$ must therefore also be horizontal, for if the graph met the line at an angle, the two would have to cross. Intuitively, then, the meeting of the line and the graph of the function is a grazing contact; the line is tangent to the graph. To locate a peak of the graph we seek it among the points where the graph has a horizontal tangent. To make some general use of this geometrical idea we express it numerically, so that it may serve as a basis for computation. Observing that the direction of the tangent can be represented numerically by its slope, we reformulate our idea: at a peak of the graph the slope of the tangent is zero. We introduce a new function $x \rightarrow f'(x)$ where $f'(x)$ is the slope of the graph of f at the point $[x,f(x)]$. If there is a peak of the graph of f at $[a,f(a)]$, then $f'(a) = 0$; to locate a peak, then, we look among the zeros of $f'(x)$. The function f' is called the derivative of f, and the slope of the tangent $f'(x)$ at $[x,f(x)]$ is called the derivative of f at x.

An Aide to the Reader

By now you may have a sense that we are very far from the point of beginning, and that you would like to know what we have accomplished. What we have done is this: we have replaced a problem about which we know very little, with a problem about which we know a great deal--to locate a peak of one function we look among the zeros of another function (the derivative). It may seem to you that the line of approach is devious, and it is still not evident that it is fruitful. We promise that it will be fruitful. You should not think that the discovery of such an avenue of investigation is beyond the powers of ordinary mortals. Whenever you become unduly impressed by the ingenuity and power of mathematical methods, reflect that an investigator will try not one but many approaches. To his admiring audience he will present the one idea that worked and never mention the failures that filled his waste basket with reams of paper. In fact, we briefly considered and rejected one idea already, that of finding the maximum value of $f(x)$ by examining a number of its values.

Before we go on to solve our best value problem, it should be said openly that the method of solution we rejected was a perfectly practical one. If the writer had not known what you are now learning about such problems, he might have proceeded by calculating values and come very close to the optimum solution*. The point is that problems of this kind arise often, and if we have a great many similar problems it pays to devote some attention to refined methods of solution. Similarly, if you wished to make just one pin, you would be content to do it by hand, but if you wished to produce pins by the million, you would put a great deal of effort into designing suitable machinery for the purpose. You will soon reach the point of view from which the solution of our present problem will appear no more consequential in the light of the methods we shall develop than the production of a single pin in the operation of a pin factory.

Turning back to our problem, we find that we have so far only replaced it by new problems. In particular, we have not clearly defined the direction of the graph at a given point and, hence, the slope of the tangent. Furthermore, even if the slope of the tangent or derivative is defined at a point, there remains the problem of describing the function f' in terms suitable for calculating the solution of the problem. To attack the problem of defining the derivative, we resort to a standard method of the calculus. The method is to determine a number by constructing a set of approximations in such a way that, so long as we allow any margin for error, we can always find an approximation to the number which is correct within the allowable error. In the language of the calculus we say the number is the limit of the set of approximations. To approximate the slope of the tangent at a point $[a,f(a)]$ of the graph of f we consider the arc of the graph between the point $[a,f(a)]$ and another point $[x,f(x)]$. To say that the tangent at $[a,f(a)]$ has a certain slope $f'(a)$ will mean now that it is possible to approximate $f'(a)$ by the slope $m(x)$ of the chord between $[x,f(x)]$ and $[a,f(a)]$. More precisely, the error in approximating $f'(a)$ by $m(x)$ can be reduced below any given tolerance by taking x close enough to a.

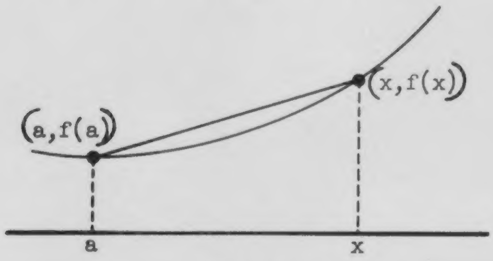


Figure 2.

---

*Since the graph is nearly horizontal in the neighborhood of a peak, the penalty for missing the exact location of the peak can be expected to be quite small. We shall return to this point later in the text.

Now we are ready to attack the box problem directly. For $f(x) = x^2(72 - 4x)$ we have

$$m(x) = \frac{f(x) - f(a)}{x - a}$$

$$= \frac{x^2(72 - 4x) - a^2(72 - 4a)}{x - a}$$

$$= \frac{72(x^2 - a^2) - 4(x^2 - a^2)}{x - a}$$

$$= \frac{x - a}{x - a}\,[72(x + a) - 4(x^2 + ax + a^2)].$$

When $x = a$ the expression for m is algebraically meaningless since the denominator is zero. This is to be expected since the geometrical interpretation of $m(x)$ as the slope of the chord joining two points loses its meaning if $[a,f(a)]$ and $[x,f(x)]$ represent the same point. We note for any other value of x than a that $\frac{x - a}{x - a} = 1$. The expression $72(x + a) - 4(x^2 + ax + a^2)$ is a polynomial which at $x = a$ has the value $144a - 12a^2$. We shall prove for a polynomial function $p(x)$ that it is possible to approximate $p(2)$ to within any fixed margin of error by taking x sufficiently close to a*. It follows that the slope of the tangent at $[a,f(a)]$ is

$$f'(a) = 144a - 12a^2.$$

Now we use the criterion that the slope of the tangent at a peak is zero. The zeros of $f'(a)$ occur at $a = 0$ and $a = 12$. The graph of f does have a horizontal tangent at $a = 0$ but, clearly, $f(0) = 0$ is not the best value. Having eliminated every other possibility, we see that the desired maximum must occur at $x = 12$. In conclusion, the largest box with length plus girth of 72 inches has square ends with 12-inch sides and a length of 24 inches.

You will have noticed that the actual computation leading to the solution is quite short. Most of the effort and time was spent in explaining the considerations underlying this method of solution. Later we shall see that we may write out $f'(a)$ on sight, and the labor of solution is then almost negligible. Finally, we have produced yet another problem to solve: if we want to find the maximum of $f(x)$ knowing the zeros of $f'(a)$, which of these, if any, yields the best value we are seeking? This is one question we shall leave to be answered in the text.

---

*This is the property we call continuity in the text.

1.2 The Concept of the Integral

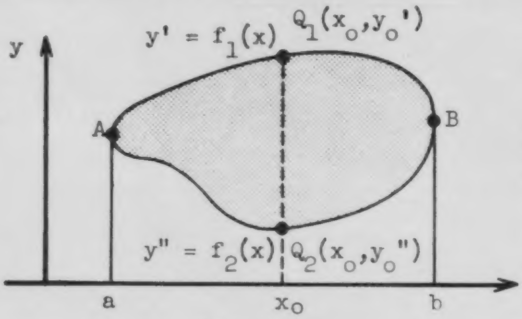


Figure 3.

The general concept of plane area is another of those geometrical ideas, like that of the direction of a curve at a given point, which remains elusive unless conceived in terms of limits. We already know a great deal about areas from geometry. We know how to calculate the areas of triangles and, hence, the areas of all figures built up of triangles, that is, the polygons. The question of determining the area of a region with curved boundaries like the shaded region in Figure 3 remains open.

You may have seen the Greek approach to the problem of determining the area of a circle, in which the area is described as the limit of areas of regular inscribed and circumscribed polygons. By using the limit of polygonal approximations, the Greeks were also able to calculate the areas of sections of a parabola, i.e., regions bounded by a line segment and a parabolic arc. In substance, the Greeks contributed nothing further to the theory of area.

In the modern theory of area we successfully make general use of the basic Greek idea of determining the area of a given region as a limit of approximations by polygonal regions. You may wonder, then, why the Greeks were not able to do so. Historians generally attribute the limitations of the Greeks in this field to their failure to develop adequate general schemes for operating with numbers. It seems that they customarily thought of real numbers in geometrical terms rather than as entities which one can study independently. Nowadays we learn to think both geometrically and numerically, taking whichever tack is the more convenient for the problem at hand. The enormous flexibility of this dual approach will enable you to solve handily problems which would have baffled the greatest Greek mathematicians.

To turn the geometrical description of the problem into a numerical one we introduce a coordinate system in the plane. For simplicity we place the axes so that the region in question is contained in the upper half plane, $y \geq 0$, as in Figure 3. Next we attack the problem of describing the region numerically. We are used to describing a curve as the graph of a function and naturally think of describing the boundary curve in terms of functions. The only difficulty is that the boundary curve is closed, which means that a vertical line will generally meet the curve more than once. In Figure 3, the vertical line $x = x_0$ meets the curve in two points, $Q_1 = (x_0, y_0')$ and $Q_2 = (x_0, y_0'')$. In fact, for this special case the boundary curve can be divided into an upper arc AQ_1B and a lower arc AQ_2B so that a vertical line intersects each arc just once. Each arc can then be considered as the graph of a non-negative function defined on the interval $a \geq x \geq b$ where a is the abscissa of A and b the abscissa of B. The numerical description of the boundary curve is now given in terms of two functions, an upper function $f_1: x \rightarrow y'$ corresponding to the arc AQ_1B and a lower function $f_2: x \rightarrow y''$ corresponding to AQ_2B. Since we have two functions to deal with, we are led to divide the calculation of the area into two parts. The area we seek is simply the difference between the areas of two regions of the same type. These are regions cut out of the strip between the vertical lines through A and B: the larger region is bounded above by the graph of f_1 and below by the x-axis; the smaller region is bounded above by the graph of f_2 and below by the x-axis.

We have reduced the problem of determining the area of the given region to the problem of determining the area of regions of a certain standard type, regions describable in terms of a single function. Of course, the region we began with was especially simple. In more complicated cases a vertical line may meet the boundary curve in more than two points and shall then need more than two functions to describe the curve. We can still approach the problem by introducing standard regions, one for each function; the method for doing so in general is left for you to think about since the details are not relevant at the moment.

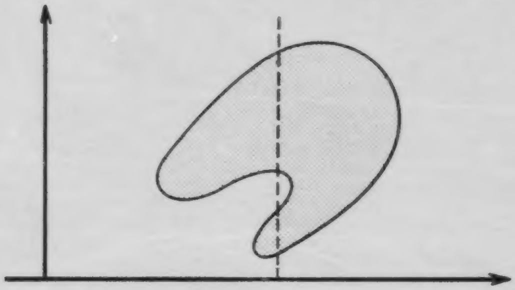


Figure 4.

We are left with the problem of calculating the area of a standard region. Given a non-negative function f on an interval $a \leq x \leq b$ we define the corresponding standard region as the set of points (x,y) for which $a \leq x \leq b$ and $0 \leq y \leq f(x)$ (Figure 5). The area of this standard region is what we call the integral of f from a to b. Again we are faced with the problem of determining a number, the area of the standard region based on the interval $a \leq x \leq b$, and the problem is apparently insoluble by any of the old methods unless the graph of f is a straight line. Again we approach the problem by treating the area as a limit. We approximate the area by polygonal areas as the Greeks did, but we are looking for a systematic scheme of approximation, one that does not depend on the particular function involved.

A first crude estimate can be given in terms of the maximum value M and

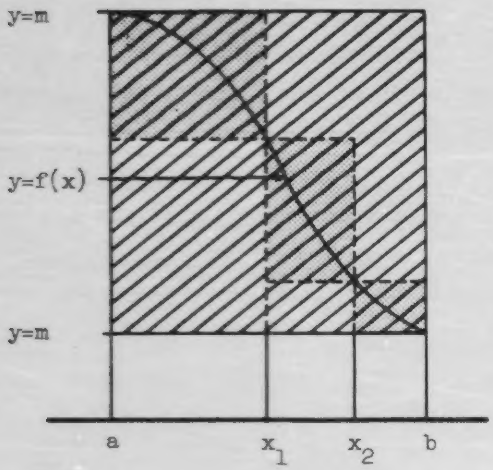


Figure 5.

the minimum value m of $f(x)$. Clearly, the rectangle of height M based on the interval $a \leq x \leq b$ contains the given region; the given region, in turn, contains the rectangle of height m on the same base. For the area A of the region we then have estimates from above and below:

$$M(b - a) \geq A \geq m(b - a).$$

If we approximate A by either of these estimates or by any value in between, then we cannot be in error by more than $(M - m)(b - a)$; that is, by the area of the hatched region indicated in Figure 5. This simple method of estimation can easily be refined in a straightforwar way. For this we only have to observe that the maximum M' of $f(x)$ on any subinterval $x_1 \leq x \leq x_2$ cannot be greater than the overall maximum M'. Similarly, the minimum m' of $f(x)$ on the subinterval cannot be less than the overall minimum m; that is,

$$m' \geq m \quad \text{and} \quad M' \leq M.$$

It follows for the area A' of the standard subregion based on the interval $x_1 < x < x_2$ that

$$m(x_2 - x_1) \leq m'(x_2 - x_1) \leq A' \leq M'(x_2 - x_1) \leq M(x_2 - x_1).$$

From this we see that the largest possible error in estimating the area of the subregion has been reduced from the former value of $(M - m)(x_2 - x_1)$ to $(M' - m')(x_2 - x_1)$. For the whole interval the maximum error can be reduced by subdividing it into smaller intervals and making the same sort of estimate separately for each of the subintervals. For the subdivision of Figure 5 this means reducing the margin of error from the area of the hatched region to that of the stippled region. Plainly, the thing to do now is to try to bring maximum error down below any given margin by making the subdivision fine enough.

The best way to see how this general approach works is to try it out on a specific function. For this purpose we try to evaluate the integral A of $f: x \rightarrow \sqrt{x}$ from 0 to 1 (Figure 6). For this function our work is simplified because the bigger the value of x, the bigger is $f(x)$. It follows that in any interval the minimum value of $f(x)$ occurs at the left endpoint and the maximum

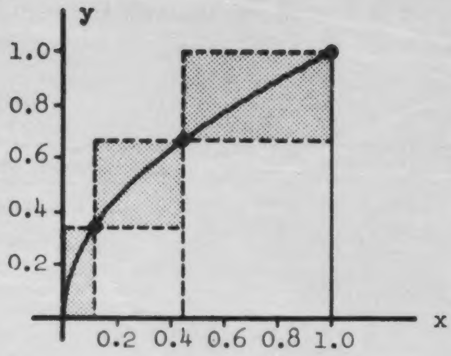


Figure 6.

value at the right. In order to calculate these values easily we choose the points of subdivision to avoid tedious calculation of the square root. If we subdivide the base into n parts we denote the successive endpoints of the n subintervals by $x_0, x_1, \ldots, x_n$ where $0 = x_0 < x_1 < \ldots < x_n = 1$. For simplicity, we choose the endpoints so that $\sqrt{x_1} = \frac{1}{n}$, $\sqrt{x_2} = \frac{2}{n}$, $\sqrt{x_3} = \frac{3}{n}, \ldots$; that is, $x_1 = (\frac{1}{n})^2$, $x_2 = (\frac{2}{n})^2$, $x_3 = (\frac{3}{n})^2$, ... Figure 6 shows the subdivision for $n = 3$. For $n = 1$, taking the entire interval, we find that the area is between 0 and 1. Taking $n = 2$, we obtain, on adding lower and upper estimates for the two intervals,

$$0 \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4} \leq A \leq \frac{1}{2} \cdot \frac{1}{4} + 1 \cdot \frac{3}{4}$$

or

$$\frac{3}{8} \leq A \leq \frac{7}{8}$$

and we have reduced the maximum error from 1 to $\frac{1}{2}$. Taking $n = 3$, we obtain similarly

$$0 \cdot \frac{1}{9} + \frac{1}{3} \cdot \frac{3}{9} + \frac{2}{3} \cdot \frac{5}{9} \leq A \leq \frac{1}{3} \cdot \frac{1}{9} + \frac{2}{3} \cdot \frac{3}{9} + 1 \cdot \frac{5}{9}$$

or

$$\frac{13}{27} \leq A \leq \frac{22}{27}$$

and the maximum error is reduced to $\frac{9}{27} = \frac{1}{3}$.

In general, for n subdivisions we have

$$A \geq \frac{0}{n}\left[\left(\frac{1}{n}\right)^2 - \left(\frac{0}{n}\right)^2\right] + \frac{1}{n}\left[\left(\frac{2}{n}\right)^2 - \left(\frac{1}{n}\right)^2\right] + \frac{2}{n}\left[\left(\frac{3}{n}\right)^2 - \left(\frac{2}{n}\right)^2\right] + \ldots + \frac{n-1}{n}\left[\left(\frac{n}{n}\right)^2 - \left(\frac{n-1}{n}\right)^2\right]$$

or

$$A \geq \frac{1}{n^3}\left\{0 \cdot 1 + 1 \cdot 3 + 2 \cdot 5 + \ldots + (n-1)(2n-1)\right\}.$$

Similarly,

$$A \leq \frac{1}{n}\left[\left(\frac{1}{n}\right)^2 - \left(\frac{0}{n}\right)^2\right] + \frac{2}{n}\left[\left(\frac{2}{n}\right)^2 - \left(\frac{1}{n}\right)^2\right] + \frac{3}{n}\left[\left(\frac{3}{n}\right)^2 - \left(\frac{2}{n}\right)^2\right] + \ldots + \frac{n}{n}\left[\left(\frac{n}{n}\right)^2 - \left(\frac{n-1}{n}\right)^2\right]$$

or

$$A \leq \frac{1}{n^3}\left\{1 \cdot 1 + 2 \cdot 3 + 3 \cdot 5 + \ldots + n(2n-1)\right\}.$$

Taking the difference between these results, we find for the maximum error E_n for this subdivision that

$$E_n \leq \frac{1}{n^3} 1 \cdot 1 + 1 \cdot 3 + 1 \cdot 5 + \dots + 1 \cdot n^2 - (n - 1)^2$$

or

$$E_n \leq \frac{1}{n^3} 1 + 3 + 5 + \dots + 2n - 1 .$$

The expression in braces is just an arithmetic progression for which we know the sum. We obtain, at last,

$$E_n = \frac{n^2}{n^3} = \frac{1}{n}.$$

For this method of subdivision, then, we can bring the error down below any given margin, simply by taking n big enough: given a margin of error a we take $n > \frac{1}{a}$.

It may seem that we have not answered the question: What is the number A? All we know is that we can approximate A to within any given margin of error. Nonetheless, in describing the integral A of f from 0 to 1 as the limit of a set of approximations we have left no room for ambiguity. We still may feel cheated. We would like to have a familiar representation for A like, say, $A = \frac{2}{3}$ (which, by the way, is just what A is in this case). Later we shall see in such simple cases how to obtain such a representation. Still, it is important to know what we cannot expect the solutions of our problems to take a familiar form: often the simplest and most comprehensible description of a number is its description as the limit of a set of approximations.

1.3 Problems Solvable Using These Concepts

The derivative and the integral are interpreted geometrically as slope and area, but these are only two among a wide range of applications and interpretations.

The derivative is associated with variation; in general, it measures rate of change. Among the many interpretations of derivative we have velocity, acceleration, electrical current, heat flow, strain, density. The integral is associated with totality; it generally measures the end result or net effect of variation. It has interpretations such as the momentum acquired by a body affected by a force, electrical charge, energy, work, volume, mass. Later we shall see that derivative and integral are complimentary ideas and that the inverse relation between them can be exploited to great advantage. The point is not the universality of the two concepts above, but that there is a calculus, a system of reckoning, which enables us to solve important problems involving these ideas and to solve them simply and quickly. Just as science enriches mathematics by providing concrete models, mathematics enriches science by providing system and organization.

To develop this calculus we have made an intuitive geometrical beginning. The intuitive approach is useful and suggestive, but eventually we need to know just how far our methods work and when they are likely to fail. For this purpose we must frame our ideas precisely and reason about them logically. We shall not attempt, however, to reduce the calculus to a complete deductive system. We shall try to label our omissions, however, so that you will be aware of the gaps to be filled if you undertake further study.

STUDENT TEXT

Chapter 1

INTRODUCTION
(second version)

We shall start out by telling you what differential calculus and integral calculus are all about. For each of these subjects there is a fundamental problem which can be formulated geometrically. The basic problem of differential calculus is that of finding lines tangent to given curves. The basic problem of integral calculus is that of finding areas of regions with curved boundaries.

We next will exhibit a simple special case of each of these general problems.

Example 1. Consider the curve (parabola) $y = x^2$.

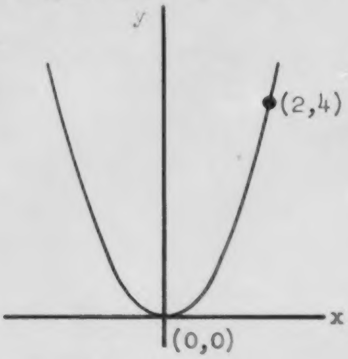


Figure 1.

We will attempt to find the equation of the line tangent to this curve at the point $P_0(2,4)$. We note that since we know one point on this line (namely, the point $(2,4)$) it will suffice to find the slope m of the tangent line. The equation of this line will then be given by

$$y - 4 = m(x - 2).$$

We are at a loss to see how to find the slope of the tangent line. One thing that we can do, however, is to choose another point $P(x,x^2)$ on the curve and find the slope of the secant line cutting the curve on the two points $P_0(2,4)$ and $P(x,x^2)$.

The slope of the line P_0P is given by

$$\frac{x^2 - 4}{x - 2} \qquad (1)$$

(regardless of whether x is greater than or less than 2).

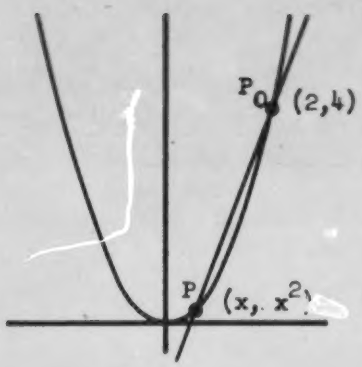


Figure 2.

Now think of the point P as moving along the curve toward P_0. The line P_0P gets closer and closer to being tangent to the curve. One might think that when P finally coincides with P_0, then the line P_0P will be tangent to the curve. Unfortunately this doesn't make any sense, as the distinct points determine a line while two coincident points do not. This is further seen by referring to the expression (1) for the slope of P_0P which becomes

$$\frac{4 - 4}{2 - 2} = \frac{0}{0} = \text{nonsense}$$

when P coincides with P_0.

It is precisely because this expression becomes nonsensical that it is necessary to introduce a new concept. And it is this new concept which distinguishes calculus from the mathematics which we have studied heretofore.

What we must do is to see what we can observe about the slope of the line segment P_0P when P is very close to P_0, that is, when x is very nearly equal to 2. Now the slope of P P, for $x \neq 2$, is, as we have seen, equal to

$$\frac{x^2 - 4}{x - 2} = \frac{(x - 2)(x + 2)}{x - 2} = x + 2. \qquad (2)$$

Now the left-hand member of (2) is not meaningful when $x = 2$, while the right-hand member is meaningful for all real numbers x. But it is only correct to say that $x + 2$ is the slope of P_0P when x is different from 2. Yet it is very easy to see that if x is close to 2 then $x + 2$ is close to 4. If P is chosen very close to P_0, then the slope of P_0P is very close to 4. We express this fact by saying "the limit as $x \to 2$ of $\frac{x^2 - 4}{x - 2}$ is 4," which we write

$$\lim_{x \to 2} \frac{x^2 - 4}{x - 2} = 4.$$

In the language of differential calculus we will say that the derivative of the function $y = x^2$ at $x = 2$ is 4. This value 4 is then the slope of the tangent line to the curve $y = x^2$ at the point $(2,4)$, and the equation of the tangent line is then

$$y - 4 = 4(x - 2).$$

Example 2. Again, look at the curve $y = x^2$.

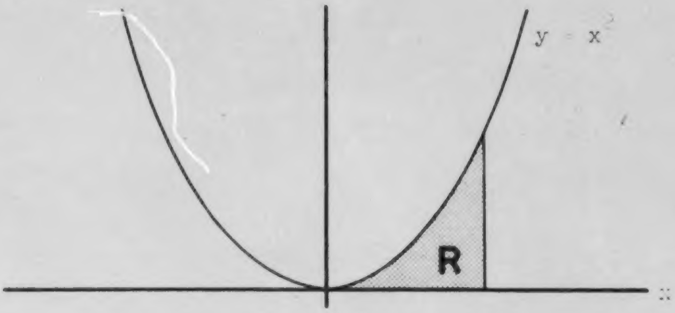


Figure 3.

We will attempt to find the area of the region R, shown in Figure 3, bounded by the curve $y = x^2$, the x-axis, and the line $x = 1$.

Although we don't know at the outset how to find this area, we do know how to find the area of a rectangle. Let us therefore attempt to approximate the area by means of regions composed of rectangles.

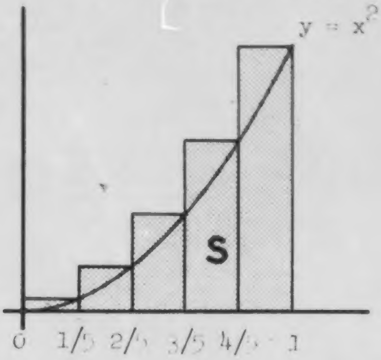


Figure 4.

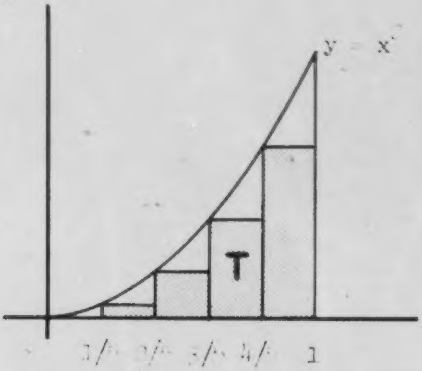


Figure 5.

The region R is contained in the shaded region S in Figure 4 and contains the shaded region T of Figure 5. Thus

$$\text{area of } T < \text{area of } R < \text{area of } S.$$

The areas of S and T are easily found as the sums of the areas of the component rectangles. In both figures the widths of the rectangles are all $\frac{1}{5}$. The heights of the rectangles of S are

$$\left(\frac{1}{5}\right)^2,\ \left(\frac{2}{5}\right)^2,\ \left(\frac{3}{5}\right)^2,\ \left(\frac{4}{5}\right)^2,\ \left(\frac{5}{5}\right)^2.$$

The heights of the rectangles of T are

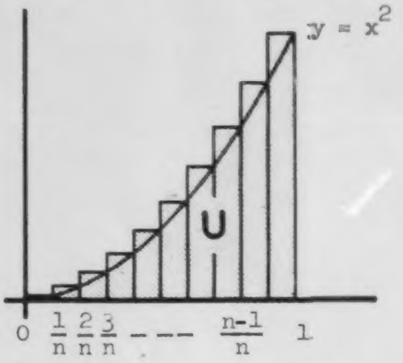
$$\left(\frac{1}{5}\right)^2,\ \left(\frac{2}{5}\right)^2,\ \left(\frac{3}{5}\right)^2,\ \left(\frac{4}{5}\right)^2.$$

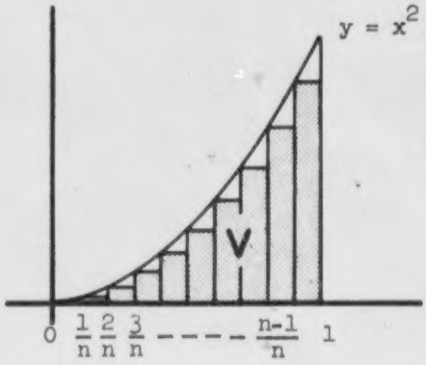
Therefore

$$\begin{aligned}\text{area of } S &= \left(\frac{1}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{2}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{3}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{4}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{5}{5}\right)^2\cdot\frac{1}{5}\\ &= \frac{1}{125}\left(1^2 + 2^2 + 3^2 + 4^2 + 5^2\right)\\ &= \frac{1}{125}\left(1 + 4 + 9 + 16 + 25\right) = \frac{55}{125} = \frac{11}{25}\\ \text{area of } T &= \left(\frac{1}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{2}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{3}{5}\right)^2\cdot\frac{1}{5} + \left(\frac{4}{5}\right)^2\cdot\frac{1}{5} = \frac{6}{25}.\end{aligned}$$

And now $\frac{6}{5} <$ area of $R < \frac{11}{25}$.

Using the same method with a larger number of rectangles gives a closer approximation of the area of R.





Here

$$\text{area of } V < \text{area of } R < \text{area of } U.$$

Again the areas of U and V are simply computed as the sums of the areas of the component rectangles. Thus

$$\begin{aligned}\text{area of } U &= \left(\frac{1}{n}\right)^2\cdot\frac{1}{n} + \left(\frac{2}{n}\right)^2\cdot\frac{1}{n} + \left(\frac{3}{n}\right)^2\cdot\frac{1}{n} + \ldots + \left(\frac{n-1}{n}\right)^2\cdot\frac{1}{n} + \left(\frac{n}{n}\right)^2\cdot\frac{1}{n}\\ &= \frac{1}{n^3}\left(1^2 + 2^2 + 3^2 + \ldots + (n-1)^2 + n^2\right)\end{aligned}$$

$$\text{area of } V = \left(\frac{1}{n}\right)^2 \cdot \frac{1}{n} + \left(\frac{2}{n}\right)^2 \cdot \frac{1}{n} + \left(\frac{3}{n}\right)^2 \cdot \frac{1}{n} + \ldots + \left(\frac{n-2}{n}\right)^2 \cdot \frac{1}{n} + \left(\frac{n-1}{n}\right)^2 \cdot \frac{1}{n}$$

$$= \text{area of } U - \frac{1}{n}.$$

It is easily shown by mathematical induction that

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Therefore

$$\text{area of } U = \frac{n(n+1)(2n+1)}{6n^3} = \frac{2n^2 + 3n + 1}{6n^2} = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}$$

and

$$\text{area of } V = (\text{area of } U) - \frac{1}{n} = \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2}.$$

Thus we find that

$$\frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} < \text{area of } R < \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}.$$

This inequality must hold for all values of the positive integer n. The right-hand and left-hand members of this inequality can be made as close to $\frac{1}{3}$ as we wish by choosing n sufficiently large. We say that the limit as n becomes large without bound of $\left(\frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}\right)$ is $\frac{1}{3}$. Similarly, the limit of $\left(\frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2}\right)$ is $\frac{1}{3}$. The number $\frac{1}{3}$ is the only number which lies between $\left(\frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2}\right)$ and $\left(\frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}\right)$ for all values of n.

It is evident then that the area of R must be equal to $\frac{1}{3}$. In the language of integral calculus we would say that the integral of x^2 from $x = 0$ to $x = 1$ is equal to $\frac{1}{3}$. This is written

$$\int_0^1 x^2 \, dx = \frac{1}{3}.$$

We have in the above seen two problems treated in a rather intuitive manner, one illustrating the methods of differential calculus, the other illustrating the method of integral calculus. In each of these problems the idea of a limit cropped up. This limit concept is the foundation of calculus.

There should now follow a list of about a dozen striking problems which can be solved by methods of calculus, for the purpose of giving the student an idea of what calculus can do.

Chapter 2

REVIEW

2.3 Functions

The effort of the scientist to understand our environment and that of the engineer to control it lead repeatedly to the attempt to determine some quantity unambiguously in terms of others. For example, an astronautical engineer who calculates the position of an orbiting satellite may fix its location if he knows the time elapsed since the launching rockets cut off, the point where cut-off occurred, and the speed and direction of motion at the instant of cut-off. To the engineer it is imperative to know that this information is sufficient to determine the position of the satellite; in other words, that there is a functional dependence of position on the other data. Examples of this kind could be multiplied endlessly, but it is clear enough from this typical instance that the elementary concept of functional dependence permeates the body of scientific thought.

A view of the idea of functional dependence may be useful to jog our memories. Loosely stated, a datum y is functionally dependent upon data $x_1, x_2, \ldots, x_n$ if each assignment of specific values to the data x_i determines y uniquely. The relation between y and the x_i is called a function, and we write

$$y = f(x_1, x_2, \ldots, x_n)$$

or, equivalently,

$$f: (x_1, x_2, \ldots, x_n) \rightarrow y$$

to indicate the functional dependence. Both expressions may be read, "f is the function which maps $(x_1, \ldots, x_n)$ onto y." Often y is referred to as the image of $(x_1, \ldots, x_n)$. For example, the area A of a triangle is functionally dependent upon the altitude h and the base b:

$$A = \frac{1}{2}bh.$$

An instructive example is the record of atmospheric pressure as a function of time plotted by a barograph at a fixed weather station. The pressure is functionally dependent upon the time since at any specific time in the historical record the pressure is uniquely determined. Clearly, functional dependence does not necessarily imply causal relation as in the satellite problem; the pressure can hardly be said to be caused by the time. Furthermore, functional dependence does not imply any "law" or "rule" like that determining the area of the triangle. There is no known rule for specifying the pressure at a given time apart from the historical record; we neither know what the atmospheric pressure was five hundred years ago nor precisely what it will be next week.

Here we shall treat only functions of the form $x \rightarrow y$ where x and y are real numbers. It is not necessary that a functional dependence be defined for all real values of x. For example, the function

$$f: x \rightarrow \sqrt{1 - x^2}$$

is defined only for x satisfying $-1 \leq x \leq 1$. The set of values of x for which the function is defined is called the domain of the function. The image of x is denoted by $f(x)$ so that we write in this particular instance

$$f(x) = \sqrt{1 - x^2}$$

and, similarly, for specific values of x we may write

$$f(1) = 0$$
$$f(-\frac{3}{5}) = \frac{4}{5}$$
$$f(0) = 1$$
etc.

The set of images $f(x)$ for x in the domain of definition of f is called the range of the function. For the function $f: x \rightarrow \sqrt{1 - x^2}$ the range consists of all values y satisfying $0 \leq y \leq 1$.

It is usually convenient to think of a function in terms of its graph, that is, the set of points (x,y) such that $y = f(x)$ (Figure 1).

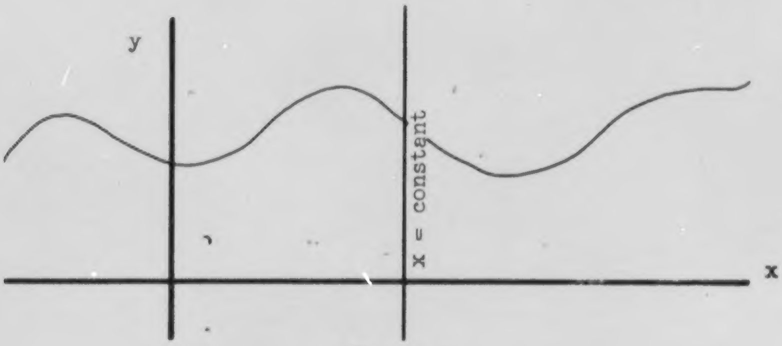


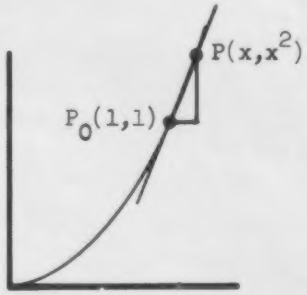
Figure 1.

The property of a function that each values of x in the domain determines just one value of y is reflected geometrically in the fact that a vertical line, x = constant, intersects the graph in no more than one point. In other words, the graph of a function is a set of points such that no two points can have the same x-coordinate. This consideration leads to the formal definition of a real function as a set of ordered pairs of real numbers such that no two pairs have the the same first member. We shall not concern ourselves with the formal definition except to note that the choice of a first member or element of the domain of the function uniquely fixes the second member or element of the range, and therefore a functional dependence is obtained under the conditions of the definition.

Chapter 3

THE DERIVATIVE

Let us return to the first example in the introduction. Here we sought to find the slope of the tangent line to the curve $y = x^2$ at the point $(1,1)$. To this end we found the value of the slope of the secant line through the points $P_0(1,1)$ and $P(x,x^2)$. This slope is seen to be $\frac{x^2 - 1}{x - 1}$. Now we let the point P approach the point P_0 along the curve. That is, we let x become very close to 1 and look at what happens to the slope of P_0P. Now



$$\text{slope of } P_0P = \frac{x^2 - 1}{x - 1} = \frac{x - 1}{x - 1}(x + 1).$$

For x not equal to 1 we see that $\frac{x - 1}{x - 1}$ is equal to 1 while, if x approaches 1, then $x + 1$ approaches 2. We therefore say that the limit as x approaches 1 of $\frac{x^2 - 1}{x - 1}$ is 2. We write this more compactly as

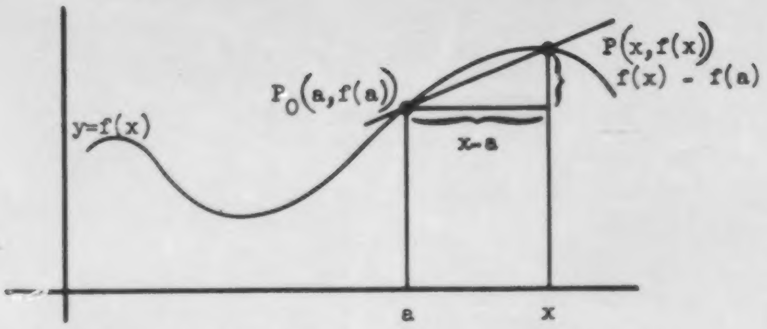
$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = 2.$$

Thus, we concluded that the slope of the line tangent to the curve $y = x^2$ at the point $(1,1)$ is 2. Since this line also passes through the point $(1,1)$, its equation is

$$y - 1 = 2(x - 1).$$

Remember that it does not make any sense to substitute 1 for x in the expression $\frac{x^2 - 1}{x - 1}$ or (what is saying the same thing) to consider the line determined by P_0 and P when these two points are coincident. What we have done instead is to investigate the limiting value of $\frac{x^2 - 1}{x - 1}$ as x approaches 1. This limiting value was easy to see when $\frac{x^2 - 1}{x - 1}$ was expressed in the form $\frac{x - 1}{x - 1}(x + 1)$. Later on you will see a precise definition of the limit concept. For the present we ask you to work with limits in the intuitive manner exhibited in this example.

Now we are ready to generalize the ideas presented in the above example. Consider a function f and try to find the slope of the line tangent to its graph at the point $P_0[a,f(a)]$. We proceed just as in the above example. Choose another point on the curve $P[x,f(x)]$ and find the slope at the secant line through these two points.



The slope of this line is given by

$$\text{slope of } P_0P = \frac{f(x) - f(a)}{x - a}.$$

Now we let P move along the curve toward P_0 or (what is saying the same thing) let x approach a. Then the limiting value of the ratio $\frac{f(x) - f(a)}{x - a}$ is the slope at the line tangent to the graph at the point $[a,f(a)]$. Note that the ratio

$$\frac{f(x) - f(a)}{x - a}$$

is meaningless when $x = a$. We cannot let $x = a$ in this ratio. We must let x approach a and find the limiting value of the ratio.

We are now ready to make the most basic definition of differential calculus.

Definition 1. The derivative of f at a written $f'(a)$ is given by

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

We shall see that the derivative arises in many other problems than that of finding tangent lines to curves.

Consider an object or particle moving on a line. Let $s(t)$ represent the displacement of the particle from a fixed point on the line, displacements on one side of the point being positive, displacements on the other side being negative. Now consider the problem of defining the velocity of the particle at a particular time t_0. We cannot see what this should be at first. It is easier to consider average velocity. The average velocity between the times t_1 and t_0 is defined as being the change in displacement divided by the elapsed time. That is,

$$\frac{s(t_2) - s(t_1)}{t_2 - t_1}.$$

The student will probably agree that to find an approximation to the instantaneous velocity at the time t_0, he would choose a time t very close to t_0 and find the average velocity between times t_0 and t,

$$\frac{s(t) - s(t_0)}{t - t_0}.$$

The closer t is to t_0, the better the approximation to the instantaneous velocity at time t_0. It should seem reasonable then to define the instantaneous velocity at time t_0 to be

$$V(t_0) = \lim_{t \to t_0} \frac{s(t) - s(t_0)}{t - t_0}.$$

Comparing this expression with the Definition 1, we see that $v(t_0)$ is nothing more nor less than the derivative $s'(t_0)$. The subject of differential calculus is devoted to the exploration of the applications and consequences of the definition of the derivative.

There follow two examples showing how to compute derivatives.

Example 1. We consider the function f defined by

$$f(x) = x^3$$

and compute $f'(2)$.

By definition

$$f'(2) = \lim_{x \to 2} \frac{f(x) - f(2)}{x - 2}$$

$$= \lim_{x \to 2} \frac{x^3 - 2^3}{x - 2}.$$

Now, recalling how to factor the difference of two cubes we see that

$$x^3 - 2^3 = (x - 2)(x^2 + 2x + 4).$$

Thus

$$f'(2) = \lim_{x \to 2} \frac{x - 2}{x - 2}(x^2 + 2x + 4).$$

Here as x approaches 2, the first factor $\frac{x - 2}{x - 2}$ remains equal to 1 while the second factor $x^2 + 2n + 4$ approaches 12. Therefore

$$f'(2) = 12.$$

Example 2. Let g be the function defined by $g(x) = \sqrt{x}$. Find $g'(3)$.

By definition

$$g'(3) = \lim_{x \to 3} \frac{g(x) - g(3)}{x - 3}$$

$$= \lim_{x \to 3} \frac{\sqrt{x} - \sqrt{3}}{x - 3}.$$

We are used to rationalizing the denominator of fractions. Here it turns out to be useful to rationalize the numerator. Thus

$$\sqrt{x} - \sqrt{3} = \frac{\sqrt{x} + \sqrt{3}}{\sqrt{x} + \sqrt{3}} \cdot \frac{\sqrt{x} - \sqrt{3}}{1} = \frac{x - 3}{\sqrt{x} + \sqrt{3}}.$$

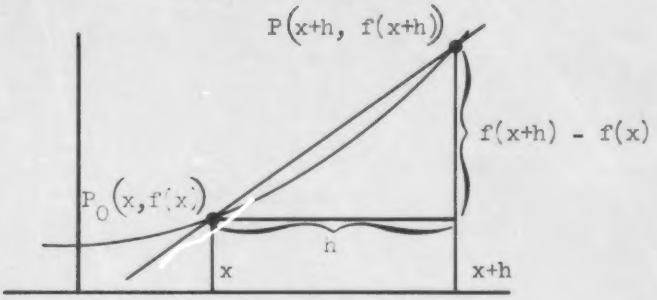
Therefore

$$g'(3) = \lim_{x \to 3} \frac{\dfrac{x - 3}{\sqrt{x} + \sqrt{3}}}{x - 3} = \lim_{x \to 3} \frac{x - 3}{x - 3} \cdot \frac{x - 3}{\sqrt{x} + \sqrt{3}}.$$

Now, as x approaches 3, the first factor $\frac{x - 3}{x - 3}$ remains equal to 1 while the second factor $\frac{1}{\sqrt{x} + \sqrt{3}}$ approaches $\frac{1}{2\sqrt{3}}$. Therefore

$$g'(3) = \frac{1}{2\sqrt{3}}.$$

A slight change in our notation makes it possible, by using the same methods as above, to find the values of $f'(x)$ for all values of x by means of a single computation. Consider a function f and consider the problem of finding the slope of the line tangent to f at the point $P_0[x, f(x)]$. Again take another point on the curve $P[x + h, f(x + h)]$ and find the slope of the secant line P_0P.



We see that

$$\text{slope of } P_0P = \frac{f(x + h) - f(x)}{h}.$$

Letting h approach 0, then P approaches P_0 and we see that the slope of the tangent line to the curve at the point $[x, f(x)]$ is given by $\lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$. We therefore have an alternate form of the definition of the derivative.

Definition 2. The derivative of f at x written $f'(x)$ is given by

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

We see from this form of the definition that differentiation (the name for the process of obtaining the derivative) yields a new function f' called the

derived function of f or simply the derivative of f. The work for Example 2, using this notation, is as follows:

Example 2a. $g(x) = \sqrt{x}$, to find $g'(x)$.

By definition

$$g'(x) = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h}$$

$$= \lim_{h \to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h}.$$

Noting that $\sqrt{x+h} - \sqrt{x} = \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}} \cdot \frac{\sqrt{x+h} - \sqrt{x}}{1} = \frac{x + h - x}{\sqrt{x+h} + \sqrt{x}} = \frac{h}{\sqrt{x+h} + \sqrt{x}}$,

we have

$$g'(x) = \lim_{h \to 0} \frac{\frac{h}{\sqrt{x+h} + \sqrt{x}}}{h} = \lim_{h \to 0} \frac{h}{h} \cdot \frac{1}{\sqrt{x+h} + \sqrt{x}}.$$

As h approaches 0, the first factor $\frac{h}{h}$ remains equal to 1 while $\frac{1}{\sqrt{x+h} + \sqrt{x}}$ approaches $\frac{1}{2\sqrt{x}}$. Thus

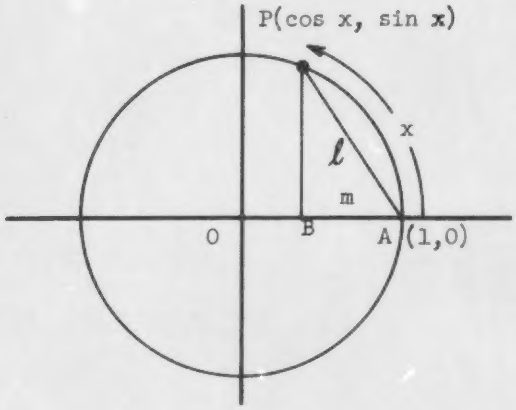
$$g'(x) = \frac{1}{2\sqrt{x}}.$$

The derivatives of the sine and cosine functions.

We will compute the derivates of the function

$$f(x) = \sin x,$$

but first it will be necessary to recall how sin x and cos x are defined. On the unit circle in the Cartesian plane (i.e., the circle with center at the origin and radius 1) measure off in a counterclockwise sense from the point A(1,0) an arc of length x.



Drop a perpendicular PB from the point P to the line OA.

The length ℓ of the segment AP is less than $|x|$ (straight line is shortest distance between two points). The length m of the segment AB is less than ℓ. (Hypotenuse of right triangle is longer than either leg.) The length m of the segment AB is equal to $1 - \cos x$. Thus

$$1 - \cos x = m < \ell < |x|.$$

Therefore

$$1 - |x| < \cos x. \tag{1}$$

Now let us compute $f'(x)$ where $f(x) = \sin x$. By definition (1)

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = \lim_{h \to 0} \frac{\sin (x + h) - \sin x}{h}.$$

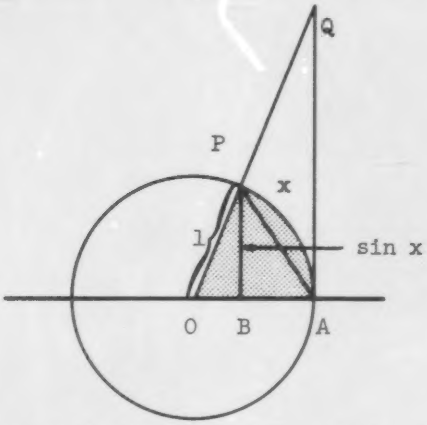
By a well-known formula of trigonometry,

$$\sin (x + h) = \sin x \cos h + \cos x \sin h.$$

Therefore,

$$\begin{aligned} f'(x) &= \lim_{h \to 0} \frac{\sin x \cos h + \cos x \sin h - \sin h}{h} \\ &= \lim_{h \to 0} [\cos x \cdot \frac{\sin h}{h} - \sin x \cdot \frac{1 - \cos h}{h}]. \end{aligned}$$

It is evident from this last expression that it is necessary to determine what happens to the ratios $\frac{\sin h}{h}$ and $\frac{1 - \cos h}{h}$ as h approaches 0.



Let O, A, P, B, X be as in the preceding figure with $0 < x < \frac{\pi}{2}$. Draw a perpendicular to OA meeting OP extended at Q. Note that the length of PB = $\sin x$, length of AQ = $\tan x$. Further note that

triangular region OAP < sector OAP < triangular region OAQ

so that

area of $\triangle$OAP < area of sector OAP < area of $\triangle$OAQ.

The areas of these three regions are seen to be

$$\text{area of } \triangle OAP = \frac{1}{2}\sin x$$
$$\text{area of sector } OAP = \frac{1}{2}x$$
$$\text{area of } \triangle OAQ = \frac{1}{2}\tan x.$$

Therefore

$$\frac{1}{2}\sin x < \frac{1}{2}x < \frac{1}{2}\tan x$$

or

$$\sin x < x < \tan x$$

where

$$1 < \frac{x}{\sin x} < \frac{1}{\cos x}$$

so that

$$1 > \frac{\sin x}{x} > \cos x$$

for $P < x < \frac{\pi}{2}$. (It should also be clear that this formula holds as well for $-\frac{\pi}{2} < x < 0$.) Recalling from (1) that $\cos x > 1 - |x|$ we have

$$1 > \frac{\sin x}{x} > 1 - |x|$$

for $0 < |x| < \frac{\pi}{2}$.

Now, $\lim\limits_{x \to 0} (1 - |x|) = 1$ so that $\frac{\sin x}{x}$, which is squeezed between 1 and $1 - |x|$, must also approach 1 as $x \to 0$. Thus

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \quad \text{or} \quad \lim_{h \to 0} \frac{\sin h}{h} = 1.$$

Recalling the formula $\sin^2 \theta = 1 - \cos \theta$, we see that $1 - \cos x = \sin^2 \frac{x}{2}$ so that

$$\frac{1 - \cos x}{x} = \frac{\sin^2 \frac{x}{2}}{x} = \frac{x}{4} \cdot \frac{\sin \frac{x}{2}}{\frac{x}{2}} \cdot \frac{\sin \frac{x}{2}}{\frac{x}{2}}.$$

In this product it is seen that as $x \to 0$ the first factor approaches 0 while each of the other factors approach 1. Therefore

$$\lim_{x \to 0} \frac{1 - \cos x}{x} = 0 \quad \text{or} \quad \lim_{h \to 0} \frac{1 - \cos h}{h} = 0.$$

Now we are ready to return to the problem of finding $f'(x)$ where $f(x) = \sin x$. We had already shown that

$$f'(x) = \lim_{h \to 0} \frac{\sin (x + h) - \sin x}{h} = \lim_{h \to 0} \left(\frac{\sin h}{h} \cos x - \frac{1 - \cos h}{h} \sin x\right).$$

Since $\frac{\sin h}{h} \to 1$ as $h \to 0$ and $\frac{1 - \cos h}{h} \to 0$ as $h \to 0$, we find that

$$f'(x) = 1 \cdot \cos h - 0 \cdot \sin x = \cos x.$$

The same methods may be employed to show that, if $g(x) = \cos x$, then $g'(x) = -\sin x$. This is left as an exercise for the student.

STUDENT TEXT

Chapter 4

LIMITS

We have seen in the previous chapter that the basic concept of the derivative was defined in terms of limits in this way:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

What is meant by a limit we left more or less to the intuition. We did give a statement that looked somewhat like a definition when we said such things as

$$"\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = 2$$

means that $\frac{x^2 - 1}{x - 1}$ gets closer and closer to 2 as x gets closer and closer to 1." The trouble with this alleged "definition" is that it involves the use of the expressions "gets" and "closer and closer", which have no precise mathematical meaning and may, in fact, mean different things to different people.

We certainly do not wish to "knock" intuition; it is in the intuition that all concepts originate. In fact, our intuitive idea of limit was adequate to handle all the limit problems met in the last chapter. After the discussion in that chapter the student should be convinced, for example, that $\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = 2$. The student ought to feel that any improved definition of limit had better yield the answer 2 for $\lim_{x \to 1} \frac{x^2 - 1}{x - 1}$ or else this definition is just no good.

The intuitive idea of limit has been quite useful and has enables us to solve a number of problems. But as we strive to extend our knowledge of the process of differentiation by deriving various rules and consequences or theorems relating to differentiation, the intuitive concept of limit becomes inadequate. What is needed is apprecise definition of limit in terms of previously encountered mathematical concepts, such as number, equality and inequality, and the logical idea of

if ..., then ... .

The proper definition is, however, not very easy to come by. In fact, for 150 years after the first great flowering of the calculus, inspired primarily by Isaac Newton, the greatest minds of the age struggled with the concept of the limit before the precise definition was finally formulated. We do not expect the student reading this text to spend as long as that in mastering the concept of limit; but neither do we expect this mastery to come overnight. We expect, based on a great deal of experience with other students, that the reader will require several years to develop complete insight into the implications of this definition and to develop the technique of using it properly to prove theorems. We emphatically do not mean to assert that the student will be unable to go on with the study of calculus until he has thoroughly mastered this definition. Analogously, we would not say that a student cannot learn to play the piano without having first mastered all the intracacies of the musical theory of harmony.

Let us return to the phrase "as x gets closer and closer to a then $f(x)$ gets closer and closer to L." This can be expressed in the form "as the distance between x and a gets very small then the distance between $f(x)$ and L gets very small," or, equivalently, "As $|x - a|$ gets very small, then $|f(x) - L|$ gets very small." The word "gets" seems to tie the whole concept up with the idea of motion from which we should like to disentangle ourselves. We might try the wording: "If $|x - a|$ is sufficiently small, then $|f(x) - L|$ is very small."

It will pay us to forget for a while the smallness of $|x - a|$ and fix our attention on the smallness of $|f(x) - L|$. How small is small? Will it help to show that $|f(x) - L| < .1$? to show that $|f(x) - L| < .01$? We recognize that smallness is relative. We can say what we mean by asserting that one number is smaller than another, but not what we mean by saying that a number is small.

It doesn't help to show that $|f(x) - L| < .1$ or that $|f(x) - L| < .01$ is less than any particular "standard of closeness."

If, however, we were able to say that $|f(x) - L|$ is less than any "standard of smallness," i.e., that for any positive number ϵ, $|f(x) - L| < \epsilon$, then surely we would be justified in saying that $|f(x) - L|$ is small. The trouble with this is that if for some value of x, $|f(x) - L| < \epsilon$ for every positive number ϵ, then for this value of x, $|f(x) - L|$ must be equal to zero. For if $|f(x) - L| = a > 0$, then it is not true that $|f(x) - L| < \frac{a}{2}$, so that $|f(x) - L| < \epsilon$ does not hold for every positive number ϵ.

Such considerations led eighteenth century mathematicians to grope for quantities called "infinitesimals" which were infinitely small, yet not zero. As we can see, such quantities could not be numbers. Today such ideas are completely rejected and relegated to the category of mysticism with the "philosopher's time" and "phlogiston." Happily there is no need for such mysticism.

Returning again to the problem of showing that

$$|f(x) - L| < \epsilon$$

for every positive number ϵ, we will find that we can actually show this if we reverse the order of doing things. That is, first we pick the positive number ϵ and then show that for certain numbers x we will have

$$|f(x) - L| < \epsilon.$$

The set of these numbers x for which this inequality holds will depend on the choice of ϵ. We do not say that there are any numbers x such that

$$|f(x) - L| < \epsilon.$$

holds for all positive numbers ϵ. We rather say that for each positive number ϵ there are numbers x for which

$$|f(x) - L| < \epsilon.$$

Now we ask, "For what values of x do we wish to be able to show that $|f(x) - L| < \epsilon$ holds?" The answer is, "for all x sufficiently close to a," or, "for all numbers x within a certain distance of a (but not equal to a)." In other words we would like to be able to show that for every positive number ϵ, there can be found a positive number δ so that for all numbers x in the interval $(a - \delta, a + \delta)$ (except for a itself), we will have $|f(x) - L| < \epsilon$. This, slightly restated, is the desired definition of $f(x) \to L$ as $x \to a$.

Definition. We say that $f(x) \to L$ as $x \to a$ if for every positive number ϵ there is a positive number δ such that, if $0 < |x - a| < \delta$, then $|f(x) - L| < \epsilon$.

Put back into the sort of language we started out with, this definition might be worded: no matter how close it might be required that $f(x)$ should be to L (within ϵ of), this degree of closeness can be guaranteed by insisting that x should be sufficiently close to a (within δ of).

It is remarkable to note that in the final definition there is no mention of smallness. What happened to it? Where did it go? The answer is that the idea of smallness is hidden in the arbitrariness of ϵ. If we can obtain a suitable δ for every positive ϵ, then we can certainly do it for small ϵ no matter what we may mean by small.

The meaning of the above definition is best appreciated if we see what it means geometrically. We wish to show that $f(x) \to L$ as $x \to a$. (We will first consider the problem without the graph of f being drawn so as to avoid any bias.) First locate a on the x-axis and L on the y-axis, as in Figure 1a.

Next choose some number $\epsilon > 0$ and draw the lines $y = L + \epsilon$ and $y = L - \epsilon$. Now we must find a number $\delta > 0$ so that for any number x in the interval

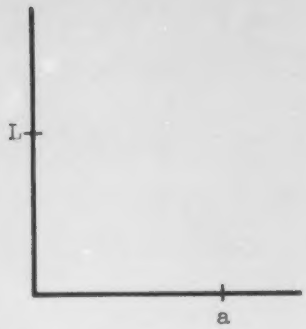


Figure 1a.

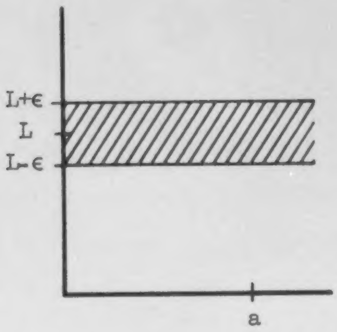


Figure 1b.

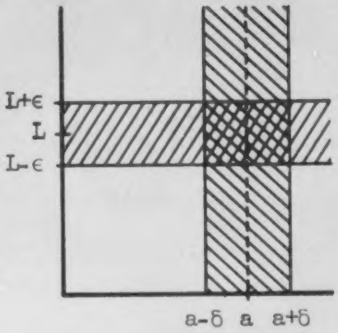


Figure 1c.

$(a - \delta, a + \delta)$ except a itself, $f(x)$ will be between $L - \epsilon$ and $L + \epsilon$, see Figure 1c. This means that for all x with $0 < |x - a| < \delta$, the point $[x, f(x)]$ of the graph of f will lie in the shaded horizontal strip of Figure 1c. Thus all points of the graph of f lying in the shaded vertical strip will also lie in the shaded horizontal strip. Next we repeat the drawings of Figures 1b and 1c with the graph of f included.

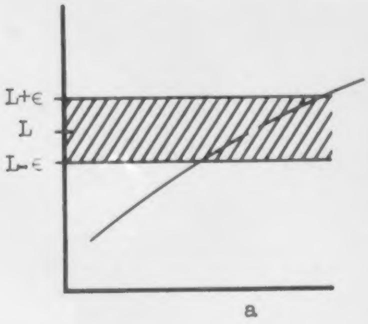


Figure 2a.

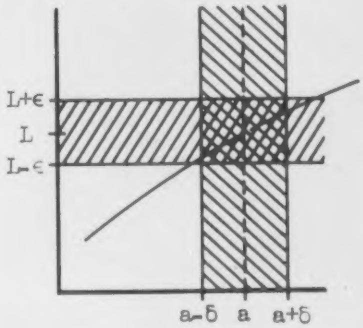


Figure 2b.

With ϵ as indicated in Figure 2a we have succeeded in finding a number δ with the required property in Figure 2b. We can see that we might have chosen δ a little larger or to be any smaller number than we actually did. We were not required to find the least possible value for δ, just a value, and that we did. We note that finding the δ for this particular ϵ does not show that $f(x) \to L$ as $x \to a$, but if we could show somehow that for every $\epsilon > 0$ there can be found a $\delta > 0$ with the desired properties, then we could be sure that $f(x) \to L$ as $x \to a$.

Before attempting anything so ambitious as showing the existence of a suitable δ for every ϵ, let us consider a problem in which we show the existence of a suitable δ for a particular ϵ.

In the problem of showing that

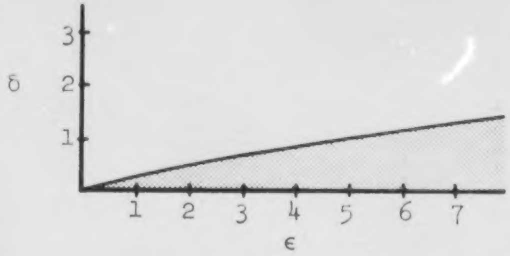
$$x^2 \to 4 \quad \text{as} \quad x \to 2$$

we would have to show that for each $\epsilon > 0$ there is a $\delta > 0$ such that

$$\text{if} \quad 0 < |x - 2| < \delta, \text{ then } |x^2 - 4| < \epsilon.$$

Let us see whether we can find a suitable δ when $\epsilon = .5$. We must find a number $\delta > 0$ such that for all numbers x differing from 2 by less than x^2 will differ from 4 by less than $.5$. Let's try a guess. Try $\delta = .1$. Now, if $0 < |x - 2| < \delta$, then $1.9 < x < 2.1$, so that $3.61 < x^2 < 4.41$, so that $-.39 < x^2 - 4 < .41$ whence $|x^2 - 4| < .41$. A lucky guess!

The student might be interested to know that for this particular problem, i.e., showing that $x^2 \to 4$ as $x \to 2$, once $\epsilon > 0$ has been chosen a suitable may be found by choosing any number satisfying $0 < \delta \leq \sqrt{\epsilon + 4} - 2$. We show this graphically



A number δ will "work" for a given ϵ if the point (ϵ, δ) lies in the shaded region. We remind the student that in these problems we never need to find the largest δ which will work for a given ϵ. This graph does illustrate two important facts, to wit:

> if a certain value of δ will work for a given value of ϵ, then any smaller (positive) value of δ will also work for this ϵ;

and

> if a certain value of δ will work for a given value of ϵ, then this value of δ will work for any larger value ϵ.

These properties are equivalent to saying that the graph of the suitable values of δ and ϵ is always the region below the graph of a monotonely increasing function.

Let us look at a particular problem and try to find a suitable δ for each value of ϵ. Consider the problem of showing that

$$\frac{1}{1 + x^2} \to \frac{1}{2} \quad \text{as} \quad x \to 1.$$

We must show that for every $\epsilon > 0$ there is a $\delta > 0$ such that

$$\text{if } 0 < |x - 1| < \delta, \text{ then } \left|\frac{1}{1 + x^2} - \frac{1}{2}\right| < \epsilon.$$

Accordingly, let $\epsilon > 0$. We cannot yet see how δ should be chosen. But for the time being we see what we can learn about what $0 < |x - 1| < \delta$ implies about $\left|\frac{1}{1 + x^2} - \frac{1}{2}\right|$ regardless of how the positive number δ may be chosen. Now,

$$\begin{aligned}
&\text{if } 0 < |x - 1| < \delta \\
&\text{then } \left|\frac{1}{1 + x^2} - \frac{1}{2}\right| = \left|\frac{2 - 1 - x^2}{2(1 + x^2)}\right| \\
&\qquad = \frac{|1 - x^2|}{2(1 + x^2)} \\
&\qquad = \frac{|1 - x|\,|1 + x|}{2(1 + x^2)} \\
&\qquad = |x - 1|\frac{|1 + x|}{2(1 + x^2)}
\end{aligned}$$

These easily followed steps are all algebraic simplifications which do not make use of the fact that $0 < |x - 1| < \delta$.

$$\begin{aligned}
&\le \delta\,\frac{|1 + x|}{2(1 + x^2)} && \text{From the fact that } |x - 1| < \delta. \\
&\le \delta\,\frac{1 + |x|}{1 + 1 + x^2 + x^2} && \text{Triangle inequality.} \\
&\le \delta\,\frac{1 + |x|}{1 + (1 + x^2)} && \text{Since size of denominator has been decreased.} \\
&\le \delta
\end{aligned}$$

The last step above follows from the fact that for all numbers x the numerator of the fraction $\frac{1 + |x|}{1 + (1 + x^2)}$ is less than the denominator, which in turn follows from the fact that for all numbers x we have $|x| < 1 + x^2$. This is obvious if $|x| \le 1$, while if $|x| > 1$ then $|x| < x^2$. Not having said anything about δ (except that it is positive), we have succeeded in showing that

$$\text{if } 0 < |x - 1| < \delta, \text{ then } \left|\frac{1}{1 + x^2} - \frac{1}{2}\right| < \delta$$

(i.e., any number x satisfying $0 < |x - 1| < \delta$ will also satisfy $\left|\frac{1}{1 + x^2} - \frac{1}{2}\right| < \delta$). If we now choose δ equal to ϵ, it will therefore be true that

$$\text{if } 0 < |x - 1| < \delta, \text{ then } \left|\frac{1}{1 + x^2} - \frac{1}{2}\right| < \epsilon.$$

LIMITS--SUPPLEMENT

Almost the only way in which one can see how the ϵ-δ definition of a limit came naturally out of the intuitive concept of limit is through a series of questions and answers which force the refinement of the ideas. We therefore present the following imaginary conversation between two high school calculus students. It took place several years ago. This explains why the dialogue is not in the most up-to-date teen-age idiom.

John and Mary are just walking down the hall after their last period calculus class. As the discussion gets more involved, they walk into an empty classroom and continue the discussion at the blackboard.

Mary: I didn't understand that limits business at all.

John: I thought I understood it pretty well. What's your problem?

Mary: Well, the teacher asked me what is the limit of x^2 as x approaches 2, and I didn't know.

John: That's easy; the answer is 4.

Mary: I see. You just substituted 2 for x in x^2 and got 4.

John: No. That's not it at all. It doesn't have anything to do with what happens when x is equal to 2. Say--wait a minute--that is how I got the answer, isn't it?

Mary: You're doing a great job of explaining this.

John: Now look, Mary, if you take that attitude you're never going to understand anything about limits.

Mary: So what?

John: You're studying differential calculus. That's all about derivatives. You're interested in derivatives, aren't you?

Mary: Wel-l-l yes, but ... .

John: Well, derivatives are defined in terms of limits, and unless you understand something about limits you can't understand derivatives--not really.

Mary: O.K. Let's try again.

John: You see, sometimes you can get the answer that way--by substituting in, but not always. For example, the limit of $\frac{\sin x}{x}$ as x approaches 0 is 1, and you can't get that one by substituting $x = 0$ in $\frac{\sin x}{x}$, can you?

180

Mary: Goodness me! If you substituted $x = 0$ in $\frac{\sin x}{x}$, you would get $\frac{0}{0}$, and everybody knows about that.

John: That's the idea. Now, when I say that the limit of x^2 is 4 as x approaches 2, I mean that for numbers x which are very close to 2, the value of x^2 will be very nearly equal to 4; that is to say, if you use 4 in place of x^2 the error will be very small.

Mary: Small? You mean like -1,000,000.

John: What? Oh, good grief, I see what you're thinking. You think that by small I mean far to the left on the number line. That isn't what I mean at all. What I mean when I say that the error made in using 4 in place of x^2 is small is that the distance between the numbers x^2 and 4 will be nearly zero. Maybe it would have been clearer if I'd said that $|x^2 - 4|$ will be small.

Mary: Yes, that is clearer, but I'm losing the gist of the argument.

John: Well, it adds up to this. What we mean when we say that the limit of x^2 is 4 as x approaches 2 is that for values of x close to 2 the value of $|x^2 - 4|$ will be small.

Mary: What's so special about 4. Won't it also be true that the value of $|x^2 - 4.0|$ will be small?

John: Oh yes, but not nearly small enough. For values of x very, very close to 2, the value of $|x^2 - 4.0|$ will be about .01, which isn't so very small.

Mary: I don't get it. What do you mean by small? How small is small anyhow?

John: I guess what I really mean is that $|x^2 - 4|$ is as small as you like. No matter how small a positive number you give me, I can show that $|x^2 - 4|$ is still smaller.

Mary: That would mean that $|x^2 - 4|$ would have to be less than every positive number.

John: That's right.

Mary: But since $|x^2 - 4|$ is the absolute value of something, it can't be less than zero.

John: That's right.

181

Mary: Well then, $|x^2 - 4|$ can only be zero because there is no positive number which is less than every positive number. But if $|x^2 - 4|$ is zero, then x is equal to 2 or -2.

John: Oh, well, you see--hm-m-m. Now ..., that is ..., how's that again?

Mary: I said .... .

John: No, I remember what you said. But, let's see ..., wel-l-l ..., gosh, you've got me all mixed up.

Mary: Goodbye, Johnny.

John: Just a minute, just a minute. I'm beginning to see it now. Don't go away.

Mary: O.K. What is it?

John: Look. Here's how it goes. When I say that the limit of x^2 is 4 as x approaches 2, I mean that the error between x^2 and 4 can be made as small as you like by requiring that x be sufficiently close to 2. That is, if you pick out a number and say that you don't want the error to exceed that number, then I have to find another numbersso that, when the distance between x and 2 is less than my number, then the error between x^2 and 4 will be less than your number.

Mary: You can do that?

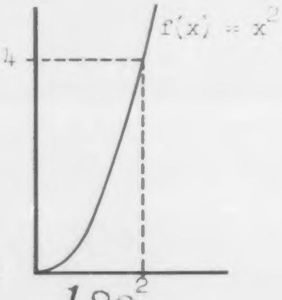
John: Try me.

Mary: .3

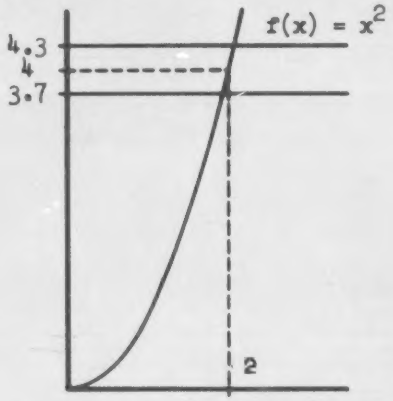
John: What?

Mary: I don't want the error to exceed .3.

John: Oh, yes. Well, let's see; that means that I have to find another number, I'll call it d, so that when the distance between x and 2 is less than d then the error will be less than .3. Maybe it would be easier to start by looking at the graph.



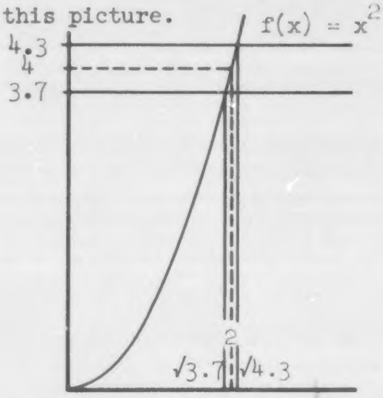
John: There is the graph of $f(x) = x^2$. The x^2 is the height of a point on the graph. You want x^2 to be within .3 of 4, so that I will draw horizontal lines at a distance .3 from the line $y = 4$ both above and below.



John: The values of x for which x^2 is within .3 of 4 will be the values of x for which the point on the graph lies between these two lines.

Mary: I'm with you so far.

John: Now, if I look at the points where these two lines cross the graph and drop projections, I get this picture.

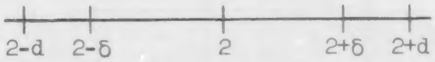


John: Now, if I let d be the smaller of the two numbers $\sqrt{4.3} - 2$ and $2 - \sqrt{3.7}$, then you can see that for all numbers x within a distance d of 2 the error between x^2 and 4 will be less than .3, as you can see from the graph.

Mary: You did it, but there's just one thing; to choose d to be the smaller of $\sqrt{4.3} - 2$ and $2 - \sqrt{3.7}$ seems rather awkward. Wouldn't it be O.K. to use a smaller but more easily written number for d?

John: That's a good idea. I didn't say that I had to find the largest possible value of d which would work. If it's true that $|x^2 - 4| < .3$ for all numbers x between $2 - d$ and $2 + d$, then it will certainly work for all numbers x between $2 - \delta$ and $2 + \delta$ when δ is smaller than d.

Mary: I see why that is true. It's because the numbers x between $2 - \delta$ and $2 + \delta$ will be a subset of the numbers x between $2 - d$ and $2 + d$. You can see that from this picture.



Anything that's true for all members of a set will also be true for all members of any subset.

John: That's what I meant but you put it much better than I did. Now let's see. What simply expressed number could I use for d? It looks like .05 ought to work. Now, $(2 + .05)^2 = 4.2025$ and $(2 - .05)^2 = 3.8025$. Both of these numbers are within .3 of the number 4, and since the function $f(x) = x^2$ is increasing, it will be true that x^2 will be within .3 of the number 4 for all x between $2 - .05$ and $2 + .05$. That means .05 is a suitable value for d.

Mary: That was quite a lot of work. Does that show that the limit of x^2 is 4 as x approaches 2?

John: Not yet. I have to be able to show that I can go through the same process no matter how small you insist that the error should be. I think that I can formulate the idea a little better now.

Mary: How?

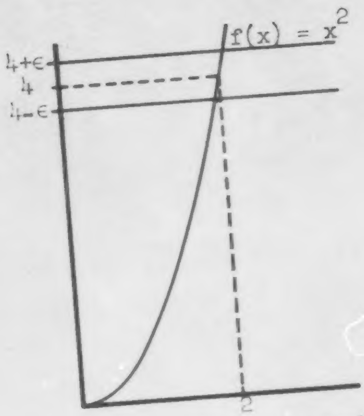
John: Well, when I say that the limit of x^2 is 4 as x approaches 2, what I mean is this: if you give me a positive number e, for error, I will be able to find a positive number d, for distance, so that for any number x whose distance from 2 is less than d it will be true that the error between x^2 and 2 will be less than e.

Mary: That's a mouthful.

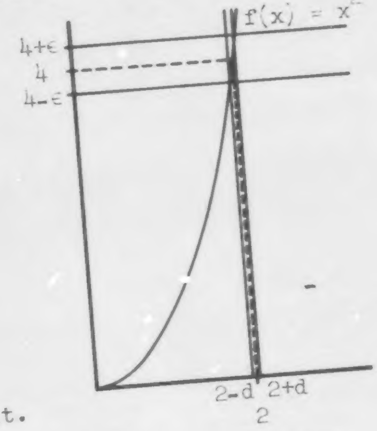
John: I can show what it means on a graph.

Mary: O.K., show me, Johnny. Well, do so by all means.

John: Now, suppose that you give me an e, I'll represent it by the length of this segment: /. Now I'll draw horizontal lines at heights $4 - \epsilon$ and $4 + \epsilon$ on the graph.



Now, I'll give you a number d; I'll represent it by the length of this tiny segment: '. Now, you can check that, for all numbers x whose distance from 2 is less than d, you'll have x^2 differing from 4 by less than e by drawing horizontal lines through 2 - d and 2 + d.



Mary: I think I've got it.

John: Well, I see the idea a lot clearer now, too.

Mary: Can you get rid of that "you give me and I'll give you" bit?

John: I'll try. By saying the limit of x^2 is 4 as x approaches 2, I mean that for every positive number e there is a positive number d so that, if the distance between x and 2 is less than d, then the error between x^2 and 4 is less than e.

Mary: I think I can clean that statement up a little bit more. Isn't it true that the distance between x and 2 is just $|x - 2|$?

John: That's right.

Mary: And the error between x^2 and 4 is just $|x^2 - 4|$?

John: Right again.

Mary: Well then, your statement can be worded: "The limit of x^2 is 4 as x approaches 2" means that for every positive number e there is a positive number d so that every number x which satisfies $|x - 2| < d$ will also satisfy $|x^2 - 4| < e$.

John: Slick! Say, you can boil it down a little more, like this: "The limit of x^2 is 4 as x approaches 2" means that for each $e > 0$ there is a $d > 0$ so that

$$\text{if } |x - 2| < d, \text{ then } |x^2 - 4| < e.$$

Mary: Does that say the same thing I said?

John: Sure. You see, after you say "for each $e > 0$ there is a $d > 0$ ...," these numbers e and d have been chosen and are fixed. The "$|x - 2| <$ " and "$|x^2 - 4| <$ " which come up afterward are conditions on x; there's nothing else for them to be conditions on; "if $|x - 2| < d$, then $|x^2 - 4| < e$" could be read as:
"If x is a number satisfying $|x - 2| < d$, then x also satisfies $|x^2 - 4| < e$."

Mary: You know, Johnny, this business of limits is coming through to me; I think I've got it. I think I see how to state the idea in general now.

John: Go ahead.

Mary: Well, the limit of $f(x)$ is L as x approaches a means that for every $e > 0$ there is a $d > 0$ so that

$$\text{if } |x - a| < d, \text{ then } |f(x) - L| < e.$$

John: You've got it. I've got it, too.

Mary: There's just one thing.

John: What?

Mary: I thought you said that the value of the limit of $f(x)$ as x approaches a doesn't have anything to do with the value of $f(a)$. Well, according to the definition we've got, if you substitute a for x in $|x - a|$, you'll have zero which is less than the positive number d, no matter what it is. That means that $|f(a) - L| < e$ for every positive number e. And that can only happen if $f(a) = L$. We've been through that before.

John: Ouch!

Mary: Does the whole thing fall apart?

John: Oh, I don't think so. All you have to do is say or remark that you don't say anything about what happens when $x = a$. Wait a minute; it's easy to do. Instead of saying

$$\text{if } |x - a| < d, \text{ then } |f(x) - L| < e,$$

we say

$$\text{if } 0 < |x - a| < d, \text{ then } |f(x) - L| < e.$$

Putting in that $0 < a$ rules at $x = a$, and it rules out nothing else. Now the definition goes like this:

The limit of $f(x)$ is L as x approaches a means: for every $e > 0$ there is a $d > 0$ so that

$$\text{if } 0 < |x - a| < d, \text{ then } |f(x) - L| < e.$$

Mary: Oh, Johnny, you've explained it to me. You're just brilliant. I could kiss you.

John: Feel free.

Mary: Some other time. I have one more question. What happened to all that business about $|x^2 - 4|$ being small?

John: Well, for gosh sakes. It's just disappeared.

Mary: Where did it go?

John: I guess that it's still there, only you don't have to talk about it.

Mary: What do you mean?

John: When you say that you can do this thing for any positive number e, this includes the small ones, no matter what you mean by "small."

Mary: I guess that's it.

STUDENT TEXT

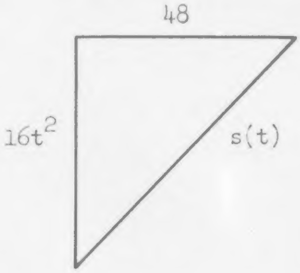
Chapter 5

THEORY AND APPLICATIONS OF THE DERIVATIVE

5.3 The Chain Rule

Example 1. A ball dropped out of a window falls $16t^2$ feet in t seconds. An observer is watching from another window at the same height 48 feet away. At what rate is the distance of the ball from the observer increasing two seconds after the ball is dropped?

To solve this problem we must first find an expression for the distance, $s(t)$, between the observer and the ball at time t. This is seen to be the hypotenuse of a right triangle with the two sides given in the data.



Thus

$$s(t) = \sqrt{48^2 + (16t^2)^2}.$$

The rate of change of $s(t)$ when $t = 2$ is $s'(2)$. The evaluation of $s'(2)$ or of $s'(t)$ from the definition of the derivative is going to be rather messy, to say the least. But $s(t)$ can be expressed in the form

$$s(t) = f(g(t))$$

where

$$g(t) = 48^2 + 256t^4$$

and

$$f(x) = \sqrt{x}.$$

We see that the function s is the composition of the functions f and g. The derivatives of f and g are easily computed by rules already learned. Thus

$$f'(x) = \frac{1}{2\sqrt{x}}$$

and

$$g'(x) = 1024t^3.$$

The question that naturally arises is: Can we somehow, through our knowledge of the derivatives of f and g, compute the derivative of the composition s of these two functions?

We can indeed, and we will proceed to do so forthwith. The theorem that gives this relationship between the derivatives of two functions and the derivative of their composition is called the "chain rule" for reasons which will presently become apparent.

Suppose we know that $g'(a)$ and $f'(g(a))$ exist. We let $s(x) = f(g(x))$, and we introduce the letter b to stand for $g(a)$ and will use b and $g(a)$ interchangeably, whichever seems most convenient. We recall that the fact $f'(b)$ exists assures us that

$$f(u) - f(b) = f'(b)(u - b) + \eta(u)(u - b) \tag{1}$$

where

$$\lim_{u\to b} \eta(u) = \eta(b) = 0. \tag{2}$$

We prefer to rewrite (1) in the form

$$f(u) - f(b) = [f'(b) + \eta(u)](u - b).$$

Now we make some simple calculations.

$$f(g(x)) - f(b) = [f'(b) + \eta(g(x))](g(x) - b).$$

Thus

$$\frac{s(x) - s(a)}{x - a} = \frac{f(g(x)) - f(g(a))}{x - a} = [f'(b) + \eta(g(x))]\,\frac{g(x) - g(a)}{x - a}$$

whence

$$\lim_{x\to a} \frac{s(x) - s(a)}{x - a} = \lim_{x\to a} [f'(b) + \eta(g(x))]\,\lim_{x\to a} \frac{g(x) - g(a)}{x - a}$$

$$= [f'(b) + \lim_{x\to a} \eta(g(x))]\,\lim_{x\to a} \frac{g(x) - g(a)}{x - a}$$

provided that these limits exist. Of course $\lim_{x\to a} \frac{g(x) - g(a)}{x - a} = g'(a)$. As for the limit $\lim_{x\to a} \eta(g(x))$, the theorem on the limit of composite functions applies here since the existence of the derivative $g'(a)$ assures us that

$$\lim_{x\to a} g(x) = g(a) = b$$

while by (2), η is continuous at b. Thus

$$\lim_{x\to a} \eta(g(x)) = \eta(b) = 0.$$

And now we find that

$$\begin{aligned} S'(a) &= \lim_{x\to a} \frac{s(x) - s(a)}{x - a} \\ &= [f'(b) + \lim_{x\to a} g(x)] \lim_{x\to a} \frac{g(x) - g(a)}{x - a} \\ &= [f'(b) + 0]\, g'(a) \\ &= f'(g(a))\, g'(a). \end{aligned}$$

Let us return to the example at the beginning of this section. Here we had found that if

$$s(t) = \sqrt{48^2 + 256t^4}$$

then

$$s(t) = f(g(t))$$

where

$$g(t) = 48^2 + 256t^4, \qquad f(x) = \sqrt{x}.$$

The problem was to find $s'(2)$. From the chain rule we learn that

$$s'(2) = f'(g(2)) \cdot g'(2).$$

Now we compute the component parts of this last expression.

$$g'(t) = 1024t^3 \quad \text{so that} \quad g'(2) = 1024 \cdot 8$$

and

$$g(2) = 48^2 + 256 \cdot 2^4 = 2304 + 4096 = 6400$$

while

$$f'(x) = \frac{1}{2\sqrt{x}} \quad \text{so that} \quad f'(g(2)) = \frac{1}{2\sqrt{6400}}.$$

Therefore

$$s'(2) = \frac{1}{2\sqrt{6400}} \cdot 1024 \cdot 8 = \frac{1}{2 \cdot 80} \cdot 1024 \cdot 8 = 51.2.$$

The answer to the problem is then that the distance of the ball from the observer is increasing at 51.2 ft/sec.

Here is another example:

Example 2. $f(x) = \sin \frac{\pi x^2}{4}$; find $f'(3)$.
Here $f(x) = g(h(x))$ where $g(y) = \sin y$ and $h(x) = \frac{\pi x^2}{4}$. Now, $g'(y) = \cos y$ and $h'(x) = \frac{\pi x}{2}$. Thus

$$f'(3) = g'(h(3)) \cdot h'(3) = (\cos \tfrac{9\pi}{4}) \cdot \tfrac{3\pi}{2} = \tfrac{3\pi}{2} \cos \tfrac{\pi}{4} = \tfrac{3\pi\sqrt{2}}{4}.$$

Substituting x for a in the conclusion of Theorem 1 yields

$$s'(x) = f'(g(x))\, g'(x).$$

We may therefore restate the theorem in the form:

Theorem 1a. If $g'(x)$ and $f'\big(g(x)\big)$ exist, then $D_x f\big(g(x)\big) = f'\big(g(x)\big)g'(x)$.

Example 3. Find $D_x(5 + 3x^2)^{27}$.

One method of solving this problem would be to expand $(5 + 3x^2)^{27}$ by means of the binomial theorem and then to differentiate term by term. But this would clearly be unpleasant. The chain rule yields a simple solution, as follows:

$$(5 + 3x^2)^{27} = f\big(g(x)\big)$$

where

$$g(x) = 5 + 3x^2 \quad \text{and} \quad f(y) = y^{27}.$$

Now

$$g'(x) = 6x \quad \text{while} \quad f'(y) = 27y^{26}.$$

Thus

$$D_x(5 + 3x^2) = 27(5 + 3x^2)^{26} \cdot 6x = 162x(5 + 3x^2)^{26}.$$

The motivation for the name "chain rule" will be seen from the following extension. Consider the problem of finding

$$D_x f\Big(g\big(h(x)\big)\Big).$$

Introduce the notation

$$s(x) = g\big(h(x)\big).$$

Then

$$D_x f\Big(g\big(h(x)\big)\Big) = D_x f\big(s(x)\big) = f'\big(s(x)\big) \cdot s'(x)$$

by means of the chain rule. But using the chain rule again

$$s'(x) = g'\big(h(x)\big) \cdot h'(x)$$

so that

$$D_x f\Big(g\big(h(x)\big)\Big) = f'\Big(g\big(h(x)\big)\Big) \cdot g'\big(h(x)\big) \cdot h'(x).$$

Let us use this formula to compute

$$D_x \sin \sqrt{1 + x^2}.$$

Here

$$\cos\,(\sqrt{1 + x^2}) = f\Big(g\big(h(x)\big)\Big)$$

where

$$f(y) = \cos y, \quad g(z) = \sqrt{z}, \quad h(x) = 1 + x^2,$$

so that

$$f'(y) = -\sin y, \quad g'(z) = \frac{1}{2\sqrt{z}}, \quad h'(x) = 2x.$$

Therefore

$$D_x \sin \sqrt{1 + x^2} = \big(-\sin\,(\sqrt{1 + x^2})\big)\,\frac{1}{2\sqrt{1 + x^2}}\,2x = -\,\frac{x}{\sqrt{1 + x^2}}\,\sin\,(\sqrt{1 + x^2}).$$

STUDENT TEXT

Chapter 6

AREA AND INTEGRAL
(revised)

Area, as we treated the idea in the introduction, was not defined but accepted as intuitively known. We did not question the idea that a region with a curved boundary has a definite numerical area but began with the implicit belief that it does, and then we proceeded to express this area in terms of a kind of limit, the integral. Considering the question again, we hardly see how to describe the area of such a region except as a limit. Given a region, our method picks out a number which we take to be the area; in effect, the method defines the area. We shall have to take this definition of area arrived at intuitively and make a statement of it in terms as precise and unambiguous as we can.

6.1 The Intuitive Concept of Area

Behind our methods of determining the area of a region, there lie a few elementary preconceived ideas about area that we never stated outright. These are the plain common-sense ideas by which people buy and sell acreage every day. The ideas are illustrated by the figures below and brought out by the following questions.

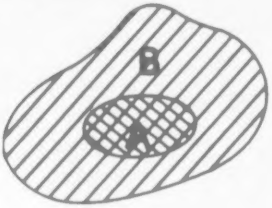


Figure 1a.

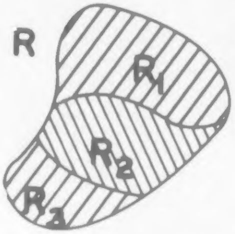


Figure 1b.

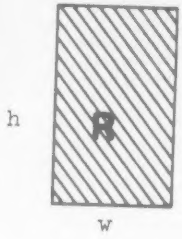


Figure 1c.

Mental Exercises

1. In Figure 1a, if the area of the doubly shaded region B is 2, then what can you say about the entire shaded region A?
2. In Figure 1b, if the areas of regions R_1, R_2, R_3 are respectively 4, 3, and 2, then what is the area of their union R?
3. In Figure 1c, if $h = 5$ and $w = 3$, then what is the area of the rectangle R?

To answer these questions, you probably used intuitively the general principles that we now write out formally.

Property 1. If A and B are two regions with $A \subset B$, then area of $A \leq$ area of B.

Property 2. If a region R is the union of several non-overlapping regions $R_1, R_2, \ldots, R_n$, then the area of R = area of R_1 + area of R_2 + ... + area of R_n.

Property 3. The area of a rectangle is the product hw of the height and the width.

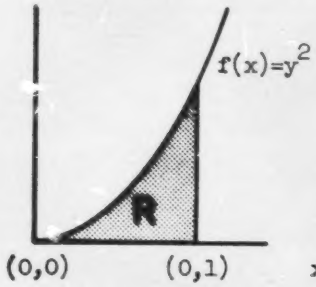


Figure 2a.

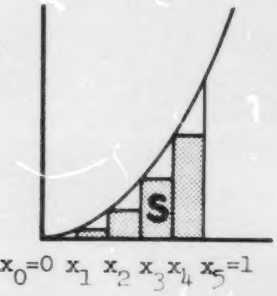


Figure 2b.

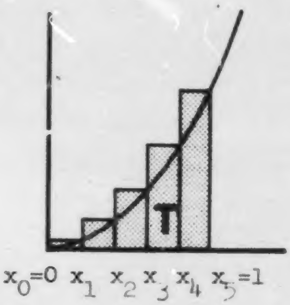


Figure 2c.

Go back to the introduction, and read over the discussion of the area problem there. You will see that these properties of area are all we have used in determining the area as a limit. There is every reason to believe that the methods of the introduction should be quite generally applicable and, in fact, we shall see that the methods using the stated properties are adequate for our purposes. These then are the properties of area which we take as basic assumptions.

Given the standard region corresponding to a non-negative function f defined on $[a,b]$, we wish to define its area as a limit. We recall that this standard region is the set of points (x,y) satisfying $a \leq x \leq b$ and $0 \leq y \leq f(x)$. For brevity, we shall refer to this as the area under $f(x)$ on $[a,b]$. If the graph of f consists of straight segments, we can find the area from geometry. We shall begin by calculating an area involving the simplest function with a curved graph, $f(x) = x^2$. As we go along, we shall point out explicitly where we use our basic assumptions about area.

Example 1. We consider the area A under $f(x) = x^2$ on $[0,1]$ shown as the dashed region R in Figure 2a. The function f is monotonically increasing, and it is easy for us to approximate A from above and below as in the introduction.

We compare R with regions S and T (Figures 2b, c) so that

$$S \subset R \subset T.$$

If C is the area of S and D is the area of T, we know from Property 1 that

$$C \leq A \leq D.$$

The regions S and T are to be made up of rectangles based on the intervals of a subdivision of $[0,1]$. In Figures 2b, c the integral $[0,1]$ is subdivided into five equal parts. In order to obtain the best possible estimates of the area by this technique, we take the height of each rectangle in S to be as large as possible without going above the part of the graph of f on its base. Similarly, the height of each rectangle in T is taken as small as possible without going below the graph of f. Now we use Property 2 to obtain

$$C = C_1 + C_2 + C_3 + C_4 + C_5$$

and

$$D = D_1 + D_2 + D_3 + D_4 + D_5$$

where C_1, C_2, C_3, C_4, C_5 are the areas of the rectangles S_1, S_2, S_3, S_4, S_5 which constitute S and D_1, D_2, D_3, D_4, D_5 the areas of the rectangles T_1, T_2, T_3, T_4, T_5 which constitute T.

The kind of repetitious writing we have just done seems labored, so we shall make use of abbreviated forms which simply indicate the repetition instead of carrying it out. Instead of writing, "S consists of the rectangles S_1, S_2, S_3, S_4, S_5," we shall write, "S consists of the rectangles S_k, $k = 1, 2, \ldots, 5$." The idea is to give the general form (S_k) in terms of an index (k) together with the values $(k = 1, 2, \ldots, 5)$ of the index for the specific items. For example, we say, "The base of S_k is the interval $x_{k-1} \leq x \leq x_k$, $k = 1, 2, \ldots, 5$," instead of making five successive statements like, "The base of S_3 is the interval $x_2 \leq x \leq x_3$." Similarly, in displaying a sum of numbers written in indexed form, it is awkward and tedious to set down each term. Instead of writing, "$C = C_1 + C_2 + C_3 + C_4 + C_5$," we shall write,

$$C = \sum_{k=1}^{5} C_k$$

Here the letter "$\sum$", sigma, which is the Greek form of our "S", stands for summation. We read, "C is the sum of C_k for k equal 1 to 5." The idea is that we take all the numbers of a general form (C_k) for the given values of the index $(k = 1, 2, \ldots, 5)$, and add them up.

Now, in our example, we choose the largest possible value for the height of

S_k, that is, the minimum of $f(x)$ on the base interval $[x_{k-1}, x_k]$ for $k = 1, 2, \dots, 5$. Furthermore, since $f: x \to x^2$ is a monotonically increasing function, we know that the minimum value is obtained at the left endpoint x_{k-1} so that the area of S_k is then given by Property 3 as

$$C_k = x_{k-1}^2(x_k - x_{k-1}) \qquad (k = 1, 2, \dots, 5).$$

For the total area of S we then have

$$C = \sum_{k=1}^{5} C_k = \sum_{k=1}^{5} x_{k-1}^2(x_k - x_{k-1}).$$

Similarly, observing that the maximum value of $f(x)$ on the interval $[x_{k-1}, x_k]$ is found at the left endpoint x_k, we have

$$D = \sum_{k=1}^{5} D_k = \sum_{k=1}^{5} x_{k-1}^2(x_k - x_{k-1}).$$

If we choose a particular subdivision of the interval $[0,1]$, that is, a set of points x_k $(k = 0, 1, \dots, 5)$ where $D = x_0 < x_1 < \dots < x_5 = 1$, then these formulas yield upper and lower estimates for the area A. For example, we might simply choose a subdivision into equal parts. In that case $x_k = \frac{k}{5}$ and

$$C_k = \frac{(k-1)^2}{25} \cdot \frac{1}{5} = \frac{(k-1)^2}{125}$$

$$D_k = \frac{k^2}{25} \cdot \frac{1}{5} = \frac{k^2}{125}.$$

Adding and factoring $\frac{1}{125}$ from the sum, we get

$$C = \frac{1}{125} \sum_{k=1}^{5} (k-1)^2 = \frac{1}{125}[0 + 1 + 4 + 9 + 16] = \frac{30}{125} = \frac{6}{25}$$

and

$$D = \frac{1}{125} \sum_{k=1}^{5} k^2 = \frac{1}{125}[1 + 4 + 9 + 16 + 25] = \frac{55}{125} = \frac{11}{25}.^*$$

Since we have chosen S and T so that $S \subset R \subset T$, we have established $C \leq A \leq D$, that is,

$$\frac{6}{25} \leq A \leq \frac{11}{25}.$$

In the absence of other information we cannot be sure of anything more than the fact that A lies in the range between the upper and lower estimates. There is no reason why A cannot be either of the end values in this integral. It follows that an approximation to A by a value A_5 somewhere between the upper and lower estimates can be in error by as much as the distance from A_5 to the farther endpoint. It seems reasonable then to choose A_5 in the middle of the integral so that the error cannot be more than half the length of the interval. With this choice of A_5 as the average of the upper and lower estimates, we then obtain

---

*You may wonder about the "rectangle" S with zero height. We'll simply agree to accept the idea of rectangles of zero height so that we don't have to consider a separate exceptional case every time this situation arises.

$$A_5 = \frac{1}{2}(C + D) = \frac{1}{2}\left(\frac{6}{25} + \frac{11}{25}\right) = \frac{17}{50} = .34$$

with an error no more than .1, that is, $|A_5 - A| < .1$.

There is nothing special about a subdivision into five parts, of course. The example is meant only to show how from the basic properties of area, Properties 1-4, we can obtain estimates of the area of the region R for any subdivision. Consequently we attempt to determine the area A of R as the limit of approximations observed by subdivision of the interval [0,1]. Using subdivisions into equal parts, we expect to be able to do better simply by dividing the integral into more parts. Without resorting to computation, it is easy to see that this is true by obtaining a bound on the error which is reduced by increasing n.

In Figures 3a, b we have subdivided the interval [0,1] into n parts by points of subdivision $x_k = \frac{k}{n}$ $(k = 0, 1, \ldots, n)$ with uniform spacing $\frac{1}{n}$. We construct regions S and T out of rectangles as before so that $S \subset R \subset T$. As before, the height of the k^{th} rectangle S_k of S $(k = 1, 2, \ldots, n)$ is taken as the minimum value of $f(x) = x^2$ on the base interval $\frac{k-1}{n} \leq x \leq \frac{k}{n}$, and the height of the k^{th} rectangle T_k of T is taken as the maximum value.

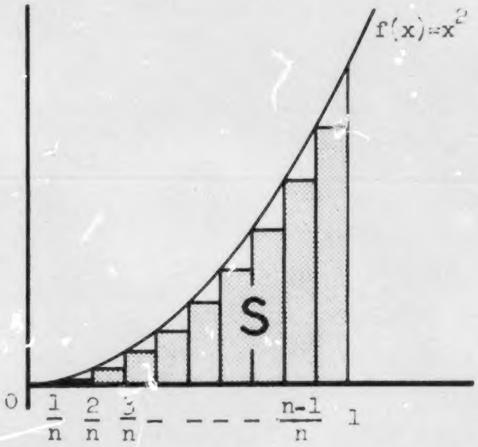


Figure 3a.

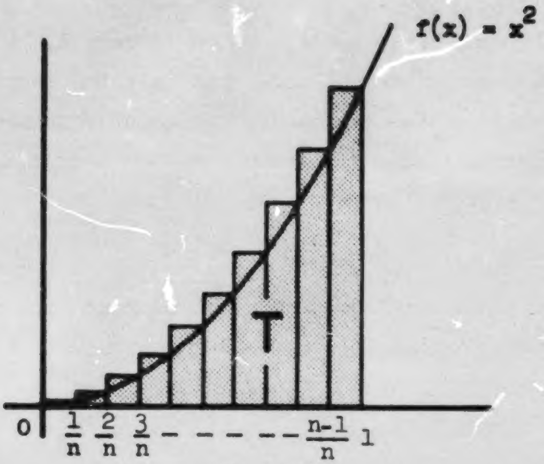
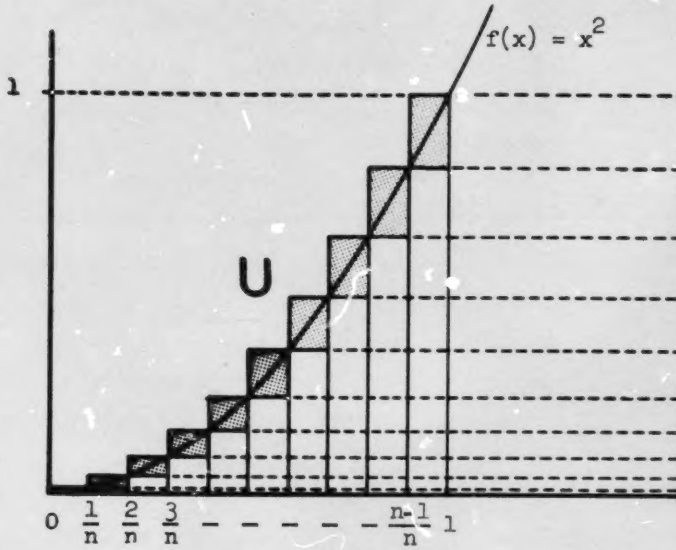

$f(x) = x^2$
T
0
$\frac{1}{n}$
$\frac{2}{n}$
$\frac{3}{n}$
$\frac{n-1}{n}$
1

Figure 3b.


$f(x) = x^2$
1
U
0
$\frac{1}{n}$
$\frac{2}{n}$
$\frac{3}{n}$
$\frac{n-1}{n}$
1

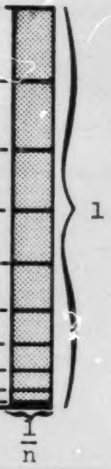


Figure 3c.

Figure 3d.

The area A of R is sandwiched between the area C of S and the area D of T, that is, $C \leq A \leq D$. The difference between C and D is the area E of the region U shown in Figure 3c. The region U consists of rectangles U_k, where U_k is the part of T_k outside of S_k. The rectangles U_k have a common base length $\frac{1}{n}$, and they can be stacked to form a rectangle of base $\frac{1}{n}$ and height 1 (Figure 3b). It follows that the total area of U is equal to the area of this rectangle, that is, $E = \frac{1}{n}$. If we estimate A by the average $A_n = \frac{1}{2}(C + D)$ of the upper and lower estimates, then, as we have already observed, the error can be no more than half the difference between these estimates, that is, $|A^* - A| \leq E$, or, for a subdivision into n equal parts,

$$|A - A_n| \leq \frac{1}{2n}.$$

At this point we know that we can find an approximation of A closer than any given margin of error. In other words, we have determined A as a limit, $A = \lim A_n$; given a margin of error ϵ we can guarantee

$$|A_n - A| < \frac{1}{2n} < \epsilon$$

simply by taking n equal subdivisions of subdivisions where $n > \frac{1}{2}$. It should be clear that we need not take A_n as the average of the upper and lower estimates to define A as a limit. We could, in fact, take any value between or even the end values themselves.

Often this is the best we can do. The method defines the area uniquely and should we need to know the area for some practical purpose we would not need to know the value exactly. Since the area can be approximated within any tolerance or error, we see that in principle the determination of the area as a limit is adequate for practical as well as theoretical purposes.

There are some nagging questions which remain to be disposed of:

If the area were a simple familiar number, would we discover it? In the preceding example it turns out that $A = \frac{1}{3}$, and our next project will be to prove that fact. The answer is that we shall show how the area can be obtained in simple form in a large variety of problems, but that it is not always clear that the area can be put in such a form even when it is possible.

Can we estimate the error better to avoid needless computation in getting a suitable approximation? We found $A_5 = .34$ above with an estimate of error given by $|A_5 - A| < .1$. Assuming $A = \frac{1}{3}$ the actual error is less than $.01$; clearly if we proposed to do some calculating, we should like to know more about such matters. We shall improve the error estimate, and in the chapter on numerical methods we shall pursue the question further.

Does the method of subdivision make any difference? We chose a very special method of subdivision, cutting the interval into equal parts. If the area were represented as a limit of approximations, under any other scheme of subdivision we ought to get the same number. This is actually so, and we shall see that it is. A much deeper question is whether the use of other kinds of approximating polygons than the kind composed of rectangular strips used here can lead to different values for the area. The answer is, "No," for the regular kind of region we consider here, but to prove it would lead us far beyond the boundaries of the present course, and such investigations are interesting primarily to the specialist.

In this general method of approximation by rectangular polygons generally applicable? The answer is that it can be used for all area problems of practical interest. We shall even prove that the method works for a sufficiently broad class of problems to cover the bulk of applications of the calculus.

In answer to the first question, we attempt to find a simple representation for the number A. The points $x_k = \frac{k}{n}$ $(k = 0, 1, \ldots, n)$ subdivide the interval $[0,1]$ into n equal parts of length $\frac{1}{n}$. The region S will be made up of the n rectangles S_k where S_k is based on the subinterval $\frac{k-1}{n} \leq x \leq \frac{k}{n}$ and has height $(\frac{k-1}{n})^2$, the minimum value of x^2 on the base interval. For the area C_k of S_k we then have $C_k = \frac{(k-1)^2}{n^3}$, and for the total area C of S

$$C = \sum_{k=1}^{n} C_k = \sum_{k=1}^{n} \frac{(k-1)^2}{n^3}$$

$$= \frac{1}{n^3}[0 + 1 + 4 + 9 + 16 + \ldots + (n-1)^2].$$

Similarly, the component rectangle T_k of T, based on the same interval $[\frac{k-1}{n}, \frac{k}{n}]$ as S_k, has height $(\frac{k}{n})^2$, the maximum value of x^2 on the base interval; the area D_k of T_k is $D_k = \frac{k^2}{n^3}$, and the total area D of T is

$$D = \sum_{k=1}^{n} D_k = \sum_{k=1}^{n} \frac{k^2}{n^3} = \frac{1}{n^3}[1 + 4 + 9 + 16 + \ldots + n^2].$$

We may verify directly that $D - C = \frac{1}{n}$, as we already knew from the picture.

The expressions for C and D are similar in form: they involve the sums of squares of consecutive integers. It would clearly be desirable to have a formula for such a sum analogous to the formula for the sum of an authentic progression. Although some mathematicians might immediately perceive a formula for the sum, the situation seems obscure to most of us. We need a trick of some kind to obtain the formula. Different people will hit upon different tricks--it is a mistake to think there is only one approach. We introduce the particular trick we use here because it is useful for much more than this special problem.

Our approach is based on the idea of a "telescoping sum." To illustrate the idea we take an excruciatingly simple example. The length of the unit interval $[0,1]$ is the sum of the lengths of the n subintervals $[\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \ldots, n$.

The sum of the lengths can be written

$$\ell = \sum_{k=1}^{n} \left(\frac{k}{n} - \frac{k - 1}{n}\right)$$

$$= \left(\frac{1}{n} - \frac{0}{n}\right) + \left(\frac{2}{n} - \frac{1}{n}\right) + \left(\frac{3}{n} - \frac{2}{n}\right) + \dots + \left(\frac{n - 1}{n} - \frac{n - 2}{n}\right) + \left(\frac{n}{n} - \frac{n - 1}{n}\right).$$

In looking at this sum we see each term added in one parentheses is subtracted in the next. The net result is that the only terms which contribute are the subtracted term in the first parentheses and the added term in the last. Consequently, $\sum_{k=1}^{n} \left(\frac{k}{n} - \frac{k - 1}{n}\right) = \frac{n}{n} - \frac{0}{n} = 1$, as we expected. In general, given numbers $u_0, u_1, \dots, u_n$, we have

$$\sum_{k=1}^{n} (u_k - u_{k-1})$$

$$= (u_1 - u_0) + (u_2 - u_1) + (u_3 - u_2) + \dots + (u_{n-1} - u_{n-2}) + (u_n - u_{n-1})$$

$$= u_n - u_0.$$

This simple idea can be used to obtain the sum of an arithmetic progression

$$\sum_{k=1}^{n} [a + (k - 1)d] = a + (a + d) + (a + 2d) + \dots + [a + (a - 1)d].$$

For this purpose we attempt to represent each term as a difference so that the sum will "telescope." The essential part of the trick is to recognize that $k^2 = (k - 1)^2 = 2k - 1$ so that we almost have k expressed in the desired form, specifically,

$$k = \frac{1}{2} [k^2 - (k - 1)^2] + \frac{1}{2}.$$

Entering this in the general form of the sum, we have

$$[a + (k - 1)d] = (a - \tfrac{1}{2}d) + \frac{1}{2} [k^2 - (k - 1)^2]d.$$

Next, we express $a - \frac{1}{2}d$ as such a difference:

$$a - \tfrac{1}{2}d = (a - \tfrac{1}{2}d)[k - (k - 1)].$$

Taking these results together, we get for the general form

$$a + (k - 1)d = u_k - u_{k-1}$$

where

$$u_k = \frac{1}{2} [k^2 + (2a - d)k]$$

and you may easily check that this is correct. In this way we represent the sum of an arithmetic progression by a telescoping formula:

$$\sum_{k=1}^{n} [a + (k - 1)d] = \sum_{k=1}^{n} u_k - u_{k-1} = u_n - u_0.$$

From the formula for u_k we get

$$u_n - u_0 = \frac{1}{2} [n^2 + (2a - d)n] - 0 = \frac{n}{2} [2a + (n - 1)d]$$

which is the old familiar formula for the sum to n terms of an arithmetic progression.

Next we turn to the problem of evaluating the sums which appear in our upper and lower estimates for the area. In both cases these involve two sums of consecutive squares, for example,

$$D = \frac{1}{n^3} \sum_{k=1}^{n} k^2 = \frac{1}{n^3} [1 + 4 + 9 + 16 + \dots + n^2].$$

To evaluate $\sum_{k=1}^{n} k^2$ we attempt to write k^2 as a difference. We recall from the preceding example that we could write k as the difference of successive squares multiplied by a constant factor plus a lower order (constant) term. This suggests attempting to write k^2 in terms of the difference of successive cubes. We have

$$k^3 - (k - 1)^3 = k^3 - (k^3 - 3k^2 + 3k - 1) = 3k^2 - 3k + 1,$$

hence,

$$k^2 = \frac{1}{3} ([k^3 - (k - 1)^3] + 3k - 1).$$

We already know how to handle the terms $3k - 1$, either as the terms of an arithmetic progression or by repeating the scheme of the preceding problem. To fix the idea we repeat the scheme. We had

$$k = \frac{1}{2} [k^2 - (k - 1)^2] + \frac{1}{2}$$

whence,

$$3k - 1 = \frac{3}{2} [k^2 - (k - 1)^2] + \frac{1}{2}.$$

Further,

$$\frac{1}{2} = \frac{1}{2} [k - (k - 1)]$$

yielding

$$3k - 1 = \frac{1}{2} [3k^2 + k] - \frac{1}{2} [3(k - 1)^2 + (k - 1)].$$

Putting all this together we get

$$k^2 = u_k - u_{k-1}$$

where

$$u_k = \frac{k^3}{3} + \frac{1}{6} (3k^2 + k).$$

We have written our sum in telescoping form:

$$\sum_{k=1}^{n} k^2 = \sum_{k=1}^{n} (u_k - u_{k-1}) = u_n - u_0 = \frac{n^3}{3} + \frac{1}{6} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}.$$

Mistakes in algebra are easily made; you will want to check this result for several values of n. (It is now plain how to proceed for the next case $\sum_{k=1}^{n} k^3$. We would attempt to express k^3 as the difference of two fourth powers; only terms of second and lower order would be left over, and the problem is then reduced to the one we have already solved. Going on in this way, we see in principle how to obtain a formula for any sum of consecutive positive integral powers.)

In the formula for the lower area estimate C, the sum is shorter by one term; hence,

$$\begin{aligned}\sum_{k=1}^{n-1} k^2 &= 1 + 4 + 9 + 16 + \dots + (n-1)^2 \\ &= [1 + 4 + 9 + 16 + \dots + (n-1)^2 + n^2] - n^2 \\ &= \sum_{k=1}^{n} k^2 - n^2 \\ &= \left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right) - n^2 \\ &= \frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6}.\end{aligned}$$

For the upper and lower estimates of area we then get

$$D = \frac{1}{n^3}\sum_{k=1}^{n} k^2 = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}$$

$$C = \frac{1}{n^3}\sum_{k=1}^{n-1} k^2 = \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2}.$$

Since the area A under $f(x) = x^2$ on $[0,1]$ is sandwiched between these values, we have

$$\frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} \leq A \leq \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}.$$

Only one number A lies between the upper and lower estimates for all n, $A = \frac{1}{3}$.

We have not only solved the problem of determining the area using only Properties 1-3, but we now see why we got such a poor estimate of the error for $n = 5$*. If we approximate A by the average of the upper and lower estimates, that is, by

---

*Of course we don't care about the error estimate in this particular problem, now that we have the exact value, but often it is impossible or inconvenient to obtain simple exact representations, and then error estimates become important.

$$A_n = \frac{1}{2}(C + D) = \frac{1}{3} + \frac{1}{6n^2},$$

then we see that

$$|A_n - A| = |A_n - \frac{1}{3}| = \frac{1}{6n^2}.$$

For the case $n = 5$, we now see that the error is $\frac{1}{150}$, where before we knew only that the error was no more than $\frac{1}{10}$.

Exercises

1. The problem posed in the introduction was to determine the area under $f(x) = \sqrt{x}$ on $[0,1]$. The summation encountered there was similar to the one encountered here. Use this fact to solve the introductory problem.
2. Obtain the result of Exercise 1 using only the fact that the area under $f(x) = x^2$ on $[0,1]$ is $\frac{1}{3}$ together with the basic assumptions about area, Properties 1-3, without further resort to summation techniques.
3. Show how the upper estimating sums for $\sqrt{x}$ are related term-by-term to the lower estimating sums for x^2.

The technique of telescoping sums used to evaluate the area under $f(x) = x^2$ is adaptable to any positive integral power and can be extended directly to any polynomial. For other functions it may not be possible to obtain simple telescoping sums, or, if it should be possible, an additional special trick may be necessary. Here is an interesting example.

Example 2. Let us attempt to find the area under the curve

$$f(x) = \cos x$$

between $x = 0$ and $x = a$ where a may be any number between 0 and $\frac{\pi}{2}$.

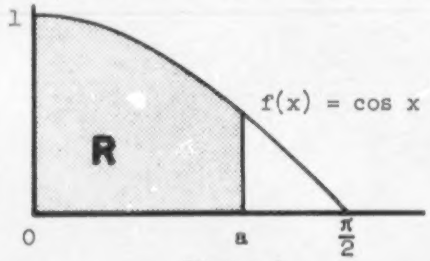


Figure 4.

We divide the interval $[0,a]$ into n subintervals of equal length and construct the containing and contained regions out of rectangles as before (Figures 5a, b).

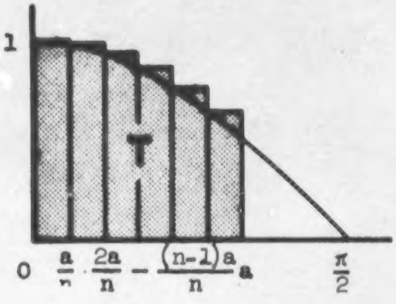


Figure 5a.

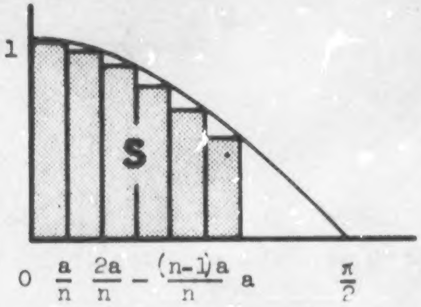


Figure 5b.

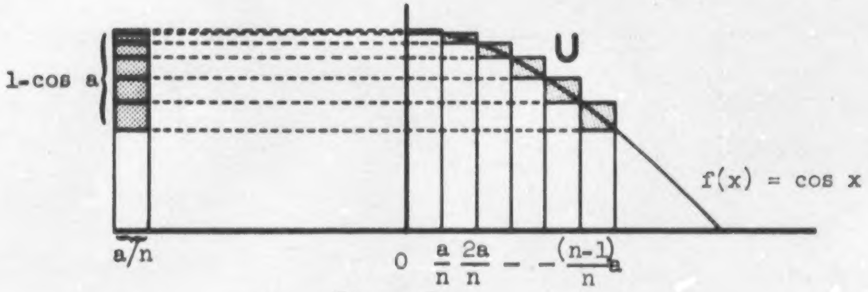


Figure 5c.

As in the preceding example, we have

$$S \subset R \subset T$$

so that

$$C < A < D$$

where C, A, and D denote the areas of S, R, and T, respectively. The set U which is left when S is taken away from T is shown in Figure 5c. From the figure we see that the area D - C of U is equal to the area of a rectangle with height 1 - cos a and base length $\frac{a}{n}$:

$$D - C = \frac{a}{n}(1 - \cos a).$$

The difference between the upper and lower estimates can therefore be reduced below any tolerance by taking n large enough.

The function f in this example is monotonically decreasing. On each interval of the subdivision f(x) will therefore take on its maximal value at the right endpoint. The rectangles S_k and T_k on the base interval $[\frac{(k-1)a}{n}, \frac{ka}{n}]$ have as their respective areas C_k and D_k where

$$C_k = \frac{a}{n}\cos\frac{ka}{n}, \quad D_k = \frac{a}{n}\cos\frac{(k-1)a}{n}.$$

Adding the areas of the component rectangles of S, we get the total area

$$C = \sum_{k=1}^{n} \frac{a}{n} \cos ka$$

$$= \frac{a}{n} \cos \frac{a}{n} + \frac{a}{n} \cos \frac{2a}{n} + \dots + \frac{a}{n} \cos \frac{na}{n}$$

$$= \frac{a}{n} [\cos \frac{a}{n} + \cos \frac{2a}{n} + \dots + \cos \frac{(n-1)a}{n} + \cos a]$$

and, similarly, for the total area D of S

$$D = \sum_{k=1}^{n} \frac{a}{n} \cos (k-1)a$$

$$= \frac{a}{n} [1 + \cos \frac{a}{n} + \cos \frac{2a}{n} + \dots + \cos \frac{(n-1)a}{n}].$$

Instead of writing $\frac{a}{n}$ repeatedly, we set $z = \frac{a}{n}$. Our problem is to rewrite as a telescoping sum

$$\sum_{k=1}^{n} \cos kz = \cos z + \cos 2z + \cos 3z + \dots + \cos nz,$$

that is, we seek values u_k $(k = 0, \dots, n)$, such that

$$u_k - u_{k-1} = \cos kz.$$

Admittedly, no one is likely to hit at once upon a suitable trick for this purpose. When the writer first solved this problem, he hit upon the following path of solution after a spell of trial and error. You might well have a more direct insight into the problem.

We wish to express a cosine as a difference. Although the cosine itself is not something one usually thinks of in that form, there is one expression involving the cosine which is often written as a difference:

$$\cos p \sin q = \frac{1}{2} [\sin (p + q) - \sin (p - q)].$$

Taking $p = kz$, we would like to fix q independently of k so that $p - q$ expressed in terms of $k - 1$ has the same form as $p + q$ expressed in terms of k. Since $p + q = kz + q$, this means that we must have $p - q = (k - 1)z + q$. On the other hand $p - q = kz - q$, so that we require $kz - q = (k - 1)z + q$. We must then have $q = \frac{1}{2}z$. In this way we find the useful result

$$\cos kz \sin \tfrac{1}{2}z = \frac{1}{2} [\sin (k + \tfrac{1}{2})z - \sin (k - \tfrac{1}{2})z]$$

$$= \frac{1}{2} [\sin (k + \tfrac{1}{2})z - \sin ([k - 1] + \tfrac{1}{2})z].$$

From the preceding result we obtain

$$\cos kz = u_k - u_{k-1}$$

where

$$u_k = \frac{\frac{1}{2}\sin(k+\frac{1}{2})z}{\sin\frac{1}{2}z}.$$

It follows that

$$\sum_{k=1}^{n}\cos kz = \sum_{k=1}^{n}(u_k - u_{k-1}) = u_n - u_0 = \frac{\frac{1}{2}\sin(n+\frac{1}{2})z}{\sin\frac{1}{2}z} - \frac{1}{2}.$$

For the area C of S we then have

$$C = z\sum_{k=1}^{n}\cos kz = \frac{\frac{1}{2}z}{\sin\frac{1}{2}z}\sin(n+\frac{1}{2})z - \frac{z}{2} = \frac{\frac{1}{2}z}{\sin\frac{1}{2}z}\sin(a+\frac{1}{2}z) - \frac{z}{2}$$

where we have set $nz = a$. There is no further difficulty in seeing what happens. We can force z to be within any given positive distance from 0 simply by taking n large enough. The limit of the expression for C as z tends to zero is easily found. We recognize $\lim\limits_{z\to 0}\frac{\frac{1}{2}z}{\sin\frac{1}{2}z}$ as the reciprocal of $\lim\limits_{h\to 0}\frac{\sin h}{h}$, which is 1. Further, since the sine function is continuous, $\lim\limits_{z\to 0}\sin(a+\frac{1}{2}z) = \sin a$. The limit of the expression for C is sin a. Since we can approximate sin a by C within any given margin of error by taking n sufficiently large, the same is true of the upper estimate D because D exceeds C only by the amount $z(1 - \cos a)$. We know that A is sandwiched between C and D. Therefore C and D can be forced to approximate both A and sin a within any given margin of error. It follows that A and sin a are the same number.

In summary, we have proved that the area under $f(x) = \cos x$ on $[0,a]$, where $0 \leq a \leq \frac{\pi}{2}$, is $A = \sin a$.

6.5 The Definite Integral

In the preceding examples we have seen how the area of a standard region can be determined as a limit of a set of approximations where the approximations are sums of areas of rectangles. In order to make general use of this scheme we outline the basic ideas.

It is important to have some idea of what is generally useful and what is just specially adapted to the solution of a particular problem. It is clear, for example, that the use of a uniformly spaced subdivision is not essential. In fact, for the example of $\sqrt{x}$ on $[0,1]$, a non-uniformly spaced subdivision proved both convenient and practical. Another special device is the use of telescoping sums. There is no evident way of applying the device of telescoping sums to finding the area under even such a simple curve as $f(x) = \frac{1}{x}$ on $[1,2]$. Later we shall see reasons why the device fails in this case, but for the while, it is enough to appreciate that it isn't always useful.

Given a non-negative continuous function f on the interval $[a,b]$, we define the standard region R under f over $[a,b]$ to be the set of points

$$R = \{(x,y): a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}.$$

We shall suppose that R has area A, and we shall attempt to approximate the number A from above and below. We divide the interval $[a,b]$ into n subintervals (not necessarily of equal length) by means of points $x_0, x_1, x_2, \ldots, x_n$ where

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$

In each of the subintervals $[x_{k-1},x_k]$ we take the maximum value M_k and minimum value m_k assumed by $f(x)$, that is,

$$M_k = \max f(x) \quad \text{for} \quad x_{k-1} \leq x \leq x_k$$

$$m_k = \min f(x) \quad \text{for} \quad x_{k-1} \leq x \leq x_k.$$

Next we take

$$\sum_{k=1}^{n} m_k(x_k - x_{k-1})$$

which is the sum of areas of non-overlapping rectangles contained in R (Figure 6a).

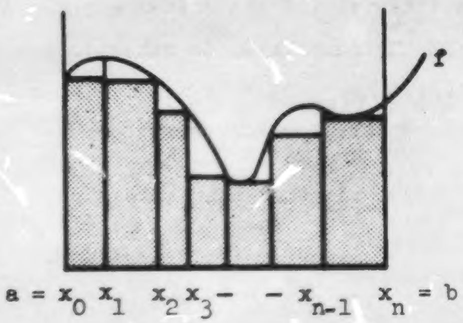


Figure 6a.

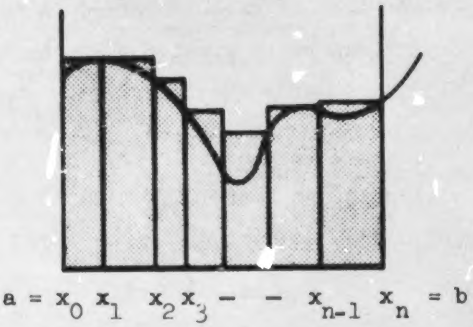


Figure 6b.

Similarly,

$$\sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

is the sum of the areas of non-overlapping rectangles which constitute a region containing R. The values of these sums are determined by the choice of particular subdivision r where

$$r = \{x_0, x_1, x_2, \ldots, x_n\}.$$

To explicitly show the dependence on r we shall write the lower sum as $L(r)$ and the upper sum as $U(r)$*:

$$L(r) = \sum_{k=1}^{n} m_k(k_k - x_{k-1})$$

$$U(r) = \sum_{k=1}^{n} M_k(x_k - x_{k-1}).$$

The area A must be sandwiched between the upper and lower sums for every choice of subdivision, that is,

$$L(r) \leq A \leq U(r).$$

Furthermore, the difference $E(r)$ between the upper and lower sums

$$E(r) = U(r) - L(r) = \sum_{k=1}^{n} (M_k - m_k)(x_k - x_{k-1})$$

gives a measure of the accuracy of approximation to A by either sum or any value between them. Specifically, if A^* is any value in the interval between $L(r)$ and $U(r)$, that is, if

$$L(r) \leq A^* \leq U(r),$$

---

*Perhaps A^- for L and A^+ for U would be a more suggestive notation.

then A and A* are contained in the same interval and the distance between A and A* can be no greater than the length of the interval. In other terms,

$$|A^* - A| \leq U(r) - L(r) = E(r).$$

Given a margin of error, the question is whether we can always find a subdivision r so that U(r) and L(r) lie closer together than this margin of error. In that case we can define A as a limit. In effect we are asking for every positive ϵ (the margin of error) that there be a subdivision r for which $E(r) < \epsilon$.

In the two preceding examples we had no trouble in finding such subdivisions. Examining these examples, we see that the essential property employed in both cases in monotonicity.

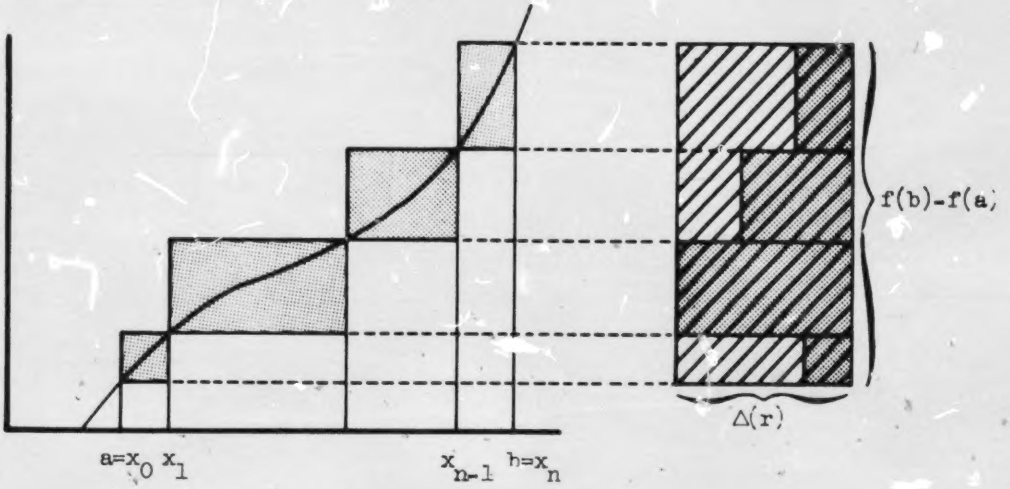


Figure 7.

The essential features can easily be appreciated geometrically. In Figure 7 we picture a monotonically increasing function f on an interval $a \leq x \leq b$ and a non-uniform subdivision of the interval. The shaded rectangle over each interval $[x_{k-1}, x_k]$ has height $M_k - m_k$ and width $x_k - x_{k-1}$; its area $(M_k - m_k)(x_k - x_{k-1})$ is therefore the k^{th} term of the sum for E(r). Because the function is monotonic, it is possible to slide these rectangles parallel to the x-axis so that the right sides all line up. In this arrangement the rectangles are contained without overlapping in a single large rectangle of height f(b) - f(a) and base length equal to the base length of the widest rectangle, that is, **max** $\{x_k - x_{k-1}\}$. For brevity, we write $\Delta(r) = \max \{x_{k-1} - x_k\}$. (The quantity $\Delta(r)$ is called the "mesh" or the "norm" of the subdivision by some authors. It is a measure of the coarseness of the subdivision.) Evidently E(r) is bounded above the area of the containing rectangle:

$$E(r) \leq [f(b) - f(a)] \max \{x_k - x_{k-1}\} = [f(b) - f(a)] \Delta(r).$$

For a monotonic function it is then quite clear how to reduce the error below any specified margin. It is only necessary to make the subdivision of the interval sufficiently fine, that is, to subdivide so that the length $\Delta(r)$ of the longest interval of the subdivision is short enough. In other terms, given any positive ϵ, we can assure that $E(r) \leq \epsilon$ by taking a subdivision so fine that

$$E(r) \leq [f(b) - f(a)] \Delta(r) < \epsilon;$$

we need only bound the length of the subintervals by

$$\Delta(r) < \frac{\epsilon}{f(b) - f(a)}.$$

We have leaned heavily on pictorial representations to obtain this result, but it is also easy to do it computationally. We have a function f continuous and monotonically increasing on $[a,b]$ and a subdivision r of $[a,b]$ where $r = \{x_k, \ k = 0, 1, 2, \ldots, n\}$ and $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. Since f is monotonically increasing, we have for the maximum M_k and the minimum m_k of $f(x)$ on $x_{k-1} \leq x \leq x_k$,

$$m_k = f(x_{k-1}) \quad \text{and} \quad M_k = f(x_k).$$

Consequently,

$$\begin{aligned} E(r) = U(r) - L(r) &= \sum_{k=1}^{n} (M_k - m_k)(x_k - x_{k-1}) \\ &= \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] (x_k - x_{k-1}). \end{aligned}$$

Since $x_k - x_{k-1} \leq \Delta(r)$, where $\Delta(r)$ is the length of the longest subinterval, and since $f(x_k) \geq f(x_{k-1})$, we conclude that

$$[f(x_k) - f(x_{k-1})] (x_k - x_{k-1}) \leq [f(x_k) - f(x_{k-1})] \Delta(r).$$

In sum we have

$$E(r) = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] (x_k - x_{k-1}) \leq \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] \Delta(r).$$

Since $\Delta(r)$ is a factor of each term of the sum, we have by the distributive law,

$$E(r) \leq \Delta(r) \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})].$$

The summation is precisely in the form of a telescoping sum, and we obtain

$$E(r) \leq \Delta(r) [f(x_n) - f(x_0)],$$

where we recall that $x_0 = a$ and $x_n = 6$. In this way we have again obtained the bound

$$E(r) \leq \Delta(r)\,[f(6) - f(a)].$$

A similar result is obtained in the same way for monotonically decreasing functions, and we leave it for you to prove.

Exercises

1. For a continuous monotonically decreasing function f on $[a,b]$, prove for the difference $E(r)$ between the upper and lower sums over the subdivision $r = \{x_k : k = 0, 1, \ldots, n\}$ where $a = x_0 < x_1 < \ldots < x_n = b$, that
$$E(r) \leq [f(a) - f(b)]\,\Delta(r),$$
where $\Delta(r) = \max\,\{x_k - x_{k-1}\}$ $(k = 1, 2, \ldots, n)$ is the length of the longest interval of the subdivision.
2. Find upper and lower estimates differing by less than .1 for the area under $f(x) = \frac{1}{x}$ on the interval $[1,2]$.

Now that we have a general method for obtaining the area under the graph of any monotone function, it is natural to want to extend the method to as large a class of functions as possible. We immediately think of those functions which can be divided into a number of monotone sections. Such functions are called "piecewise monotone." More precisely, a function f is "piecewise monotone" on $[a,b]$ is there exists a subdivision of $[a,b]$ such that f is monotone on each of the intervals of the subdivision.

Now suppose that f is piecewise monotone on $[a,b]$ and that p is the number of monotone "pieces" of f, that is, the number of intervals of a subdivision in which f is monotone on each subinterval. Let the maximum and minimum values of $f(x)$ on the interval $[a,b]$ be denoted by M and m, respectively. Then, for any subdivision r of $[a,b]$ we have

$$U(r) - L(r) \leq p(M - m)\,\Delta(r).$$

This statement is easily understood geometrically (Figures 8, 9). Figure 8a shows a function f; Figure 8b shows how the interval $[a,b]$ may be subdivided into three pieces on each of which f is monotone.

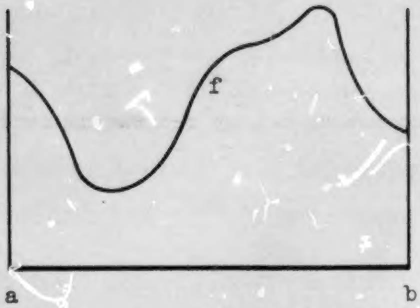


Figure 8a.

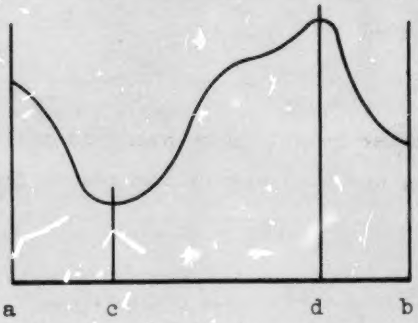


Figure 8b.

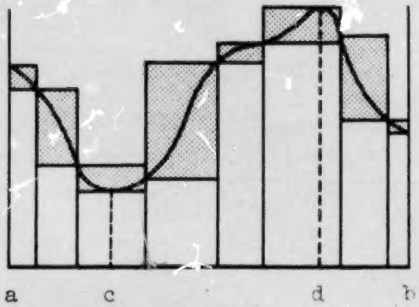


Figure 9a.

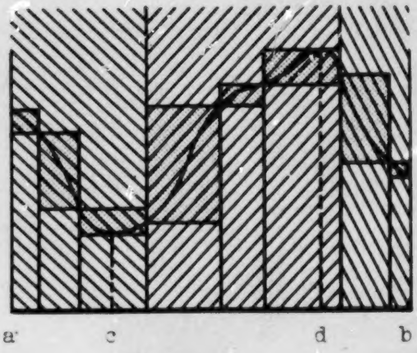


Figure 9b.

In Figure 9b we have added a partition and shaded the area corresponding to $E(r) = L(r) - U(r)$. In Figure 9b we see how to split the rectangles of Figure 9a into p groups $(p = 3)$. It is clear that the total area of the rectangles in each group is less than $(M - m)\,\Delta(r)$ since for each group the projections of the rectangles on the y-axis do not overlap! Thus since there are p such groups in all,

$$E(r) \leq p(M - m)\,\Delta(r).$$

Again we see that by making the subdivision sufficiently fine, we can approximate the area to within any given margin of error. Specifically, for each positive ϵ we can assume that

$$E(r) \leq p(M - m)\,\Delta(r) < \epsilon$$

simply by requiring

$$\Delta(r) < \frac{\epsilon}{p(M - m)}.$$

We have seen that for continuous piecewise monotone functions it is possible to define the area under the graph as a limit. The question arises whether it is possible to extend this result to an even larger class of functions. We may ask if it is possible to obtain the result for all continuous functions. Unfortunately, it is hard to visualize a typical continuous function. The graph in Figure 10 depicts an unusually tame member of the breed. For example, the limitations of the pencil might lead you to think that every continuous function is piecewise monotone but, in fact, there may be no interval anywhere, no matter how small, in which the function is monotone. If this has served to make you wary of basing everything on geometrical intuition in discussing continuous functions, so much the better. The point is that the analytical approach based on number is necessary if we are to put any faith in our conclusions. For example, we have, in effect, assumed that there is a unique number A, the area under the graph of f on $[a,b]$. Yet in contemplating the possible extravagances of an arbitrary continuous function, we are led to wonder.

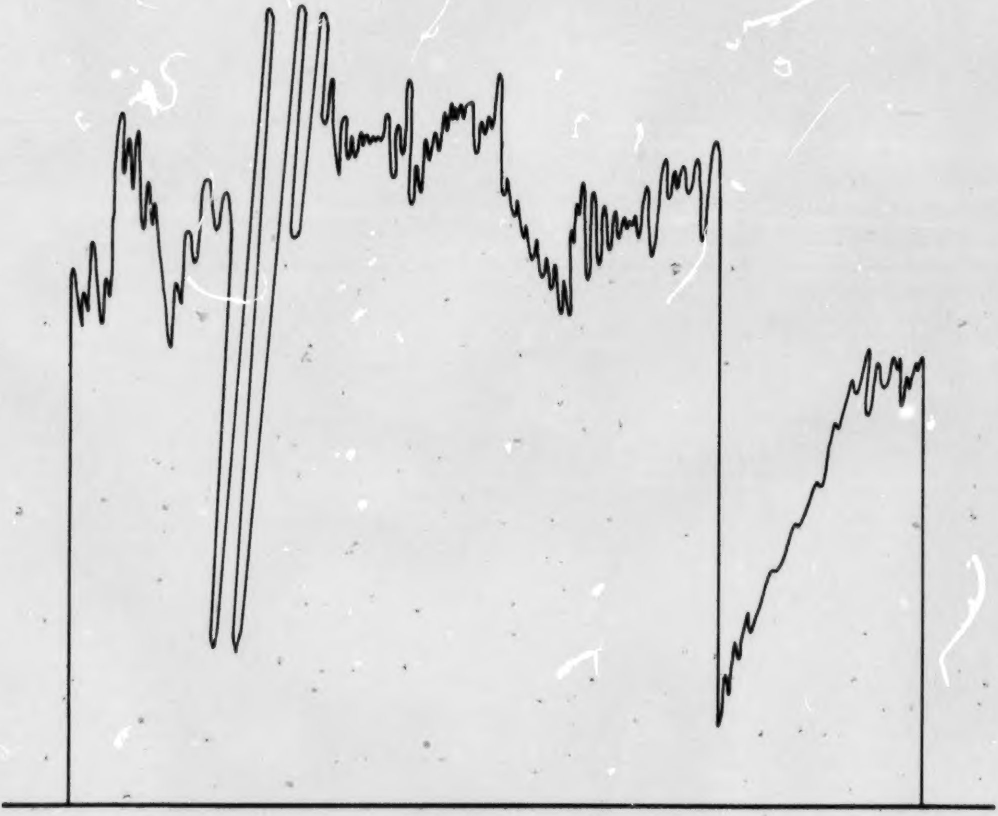


Figure 10.

We assumed that there is an area A under the graph of f on $[a,b]$ and took as our task the problem of calculating it. Given any subdivision r of [a,b], we then concluded under the basic postulates about area, Properties 1-[illegible], that

$$L(r) \leq A \leq U(r).$$

Consequently, given any pair of subdivisions r_1 and r_2, we have

$$L(r_1) \leq A \leq U(r_2),$$

that is, every lower sum is less than every upper sum. This seems intuitively reasonable, but what happens if we leave out our assumption that a number A having this property exists? We have no intuition about continuous functions in general, and we do not want to make this assumption; we want to prove that the area exists. Clearly we shall have to prove the result analytically.

Suppose that f is a function on $[a,b]$ and that r is a partition of this interval. Let us see what happens to the upper and lower sums when one additional point, x', is adjoined to this subdivision between x_{j-1} and x_j, to form a new subdivision r'.

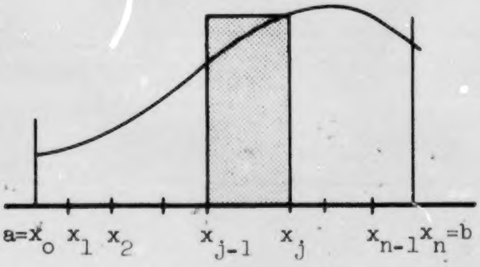


Figure 11a.

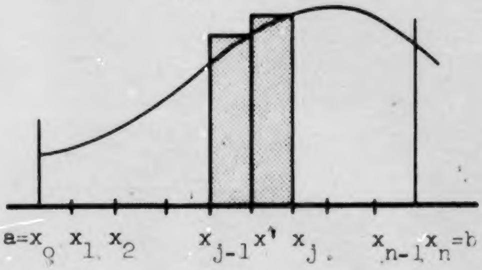


Figure 11b.

We see that all but one of the terms of

$$U(r) = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

appear also in the sum $U(r')$. The exceptional term of $U(r)$

$$M_j(x_j - x_{j-1})$$

is replaced by

$$M_j'(x' - x_{j-1}) + M_j''(x_j - x')$$

where

$$M_j' = \max f(x) \quad \text{for} \quad x_{j-1} \leq x \leq x'$$

and

$$M_j'' = \max f(x) \quad \text{for} \quad x' \leq x \leq x_j.$$

(See Figures 14a and 14b.)

Since the maximum value of a function on a subinterval cannot be larger than the maximum on the whole interval

$$M_j' \leq M_j \quad \text{and} \quad M_j'' \leq M_j,$$

so that

$$M_j'(x' - x_{j-1}) + M_j''(x_j - x') \leq M_j(x' - x_{j-1}) + M_j(x_j - x') = M_j(x_j - x_{j-1}).$$

This inequality shows us that the sum of the terms replacing $M_j(x_j - x_{j-1})$ in obtaining $U(r')$ from $U(r)$ cannot be larger than the term they replaced. Therefore

$$U(r') \leq U(r).$$

In just the same way it can be shown that $m_j' \leq m_j$ and $m_j'' \geq m_j$, so that

$$L(r') \geq L(r).$$

Now we are ready to see why, for any two subdivisions r_1 and r_2 of $[a,b]$, we must have

$$L(r_1) \leq U(r_2).$$

For, let r_3 be the subdivision consisting of all the points of both r_1 and r_2. Now, r_3 can be obtained from r_2 by starting with r_2 and successively adjoining one point of r_1 at a time. At no stage of this process is the upper sum ever increased. Thus the final upper sum in this process, $U(r_3)$, cannot be greater than the initial one, $U(r_2)$. That is,

$$U(r_3) \leq U(r_2).$$

In this way we see that

$$L(r_1) \leq L(r_3).$$

It is, of course, obvious that

$$L(r_3) \leq U(r_3)$$

since the terms in the sum $L(r_3)$ are term by term less than or equal to those in $U(r_3)$. In conclusion, we have

$$L(r_1) \leq L(r_3) \leq U(r_3) \leq U(r_2).$$

The result we have just obtained can be formulated as follows. Suppose that we have a function f, defined on $[a,b]$. Suppose we let $\mathcal{L}$ represent the set of all lower sums and $\mathcal{U}$ represent the set of all upper sums. Then every member of $\mathcal{L}$ is less than or equal to every member of $\mathcal{U}$.

When this result is formulated in this way, we see that we can assert, from Property 8 of the real numbers, that there is at least one number A separating the two sets $\mathcal{L}$ and $\mathcal{U}$. That is, for any members $U(r_1)$ of $\mathcal{L}$ and $U(r_2)$ of $\mathcal{U}$, we will have

$$L(r_1) \leq A \leq L(r_2).$$

If we were able to show that there is just one such number S which separates $\mathcal{L}$ and $\mathcal{U}$, then we would have just what we want. Under the basic assumptions, Properties 1-3, there is just one value which the area under a non-negative f on the interval $[a,b]$ could possibly have. Referring again to Chapter 2, we see that in order to demonstrate the uniqueness of the number S separating and , we have only to show that for every positive number ϵ there exist members $L(r_1)$ of $\mathcal{L}$ and $U(r_2)$ of $\mathcal{U}$, such that

$$U(r_2) - L(r_1) < \epsilon.$$

It is interesting to note that, when f is monotone, a considerably stronger result has already been demonstrated: if f is monotone in $[a,b]$, then for any subdivision r of $[a,b]$, we have

$$U(r) - L(r) < |f(b) - f(a)| \, \Delta(r).$$

Since $|f(b) - f(a)|$ is fixed, we see that by making $\Delta(r)$ sufficiently small we can force $U(r) - L(r)$ to be as small as we wish. A similar result holds for piecewise monotone functions in view of the inequality

$$U(r) - L(r) \leq p(M - m) \, \Delta(r)$$

where p is the number of intervals into which $[a,b]$ must be subdivided in order that f should be monotone on the subintervals, and where M and m are the overall maximum and minimum of $f(x)$ on $[a,b]$. Again M, m, and p are fixed, so that by choosing $\Delta(r)$ sufficiently small we may make $U(r) - L(r)$ as small as we like.

The situation is quite different if we know only that the function f is continuous. In the above discussion we found for every function which is monotone or piecewise monotone on an interval $[a,b]$ that we can find a number k so that

$$U(r) - L(r) < k \, \Delta(r).$$

It is not always true for continuous functions, however, that such a k can be found. Nevertheless, it is true that for any fucntion continuous on $[a,b]$ we can make $L(r) - U(r)$ as small as we like by choosing $\Delta(r)$ sufficiently small:

Theorem 1. If f is continuous on $[a,b]$, then for every $\epsilon > 0$, there is a $\delta > 0$ such that $U(r) - L(r) < \epsilon$ whenever $\Delta(r) < \delta$.

We will not prove this theorem in this text, for there are additional technical difficulties in our way. The student should recognize that we have proved the theorem for continuous functions which are piecewise monotone. We shall, however, assume the theorem to be true for all continuous functions. From this assumption we see that it follows that for every continuous function there is a unique number S which separates the sets $\mathcal{L}$ and $\mathcal{U}$. Thus we make the following definition:

Definition 1. If f is continuous on $[a,b]$, we define

$$\int_a^b f(x)\,dx,$$

called the integral of f over the interval $[a,b]$, to be the unique number A such that

$$U(r_1) \leq A \leq U(r_2)$$

whenever r_1 and r_2 are subdivisions of $[a,b]$.

Letting $\epsilon > 0$ and choosing $\delta > 0$ such that $U(r) - L(r) < \epsilon$ whenever $\Delta(r) < \delta$, and recalling that $L(r) < \int_a^b f(x)\,dx < U(r)$, we see that both the upper sum and the lower sum will differ from the integral by less than ϵ. Thus if the subdivision is sufficiently fine, the upper and lower sums will be close approximations of the integral. Notice, incidentally, that the definition of integral is completely numerical. In particular, the definition doesn't exclude the possibility that $f(x)$ may have negative values. The concept of integral is, therefore, not identical with that of area unless f is a non-negative function.

We re-emphasize that in order for this definition to make sense it is necessary to have Theorem 1 at our disposal to ensure the uniqueness of the number A. We should not forget that this theorem has not been proved in this book under the hypothesis that f is continuous. On the other hand, we re-emphasize that we have proved the theorem under the additional hypothesis that f is piecewise monotone. All the functions that we will encounter in this course (except for a small number of horrible examples) will be piecewise monotone. Thus, for the functions we will actually use, the theorem on which this definition rests has been proved. We state the definition for continuous functions only to bow to tradition.

A remark on the notation involed in

$$\int_a^b f(x)\,dx$$

is in order. It should be clear that the value of this integral is synonymous with the area under the function f over the interval $[a,b]$ and depends only on the function f and the interval $[a,b]$. In particular, x has nothing to do with it whatsoever. Thus

$$\int_a^b f(x)\,dx \qquad \int_a^b f(y)\,dy \qquad \int_a^b f(t)\,dt$$

all have exactly the same meaning. For that reason some authors instead write

$$\int_a^b f.$$

We will, however, adhere to the more standard notation

$$\int_a^b f(x)\,dx$$

which, in fact, has a number of advantages which will become apparent shortly. One of these advantages is that this notation permits us to write

$$\text{"}\int_a^b x^2\,dx\text{"}$$

instead of

$$\text{"}\int_a^b f$$

where f is defined as

$$f(x) = x^2.\text{"}$$

STUDENT TEXT

Chapter 6

AREA AND INTEGRAL
(first draft)

Now we are ready to study the second fundamental problem of calculus discussed in the introduction of this book, that of finding areas of regions with curved boundaries. Actually the problem is not really to find the areas of such regions but rather to give a definition of areas of such sets. So far area has only been defined for polygonal regions. However, we are not entirely free to define the areas of such regions in any way we please. We must make the definition in such a way that area will behave in accordance with our preconceived notions. We will discover, in fact, that if we adopt as assumptions a small number of these intuitive ideas, then the possible values we could adopt as the area of any particular region is narrowed down to a single value.

These intuitive ideas are exemplified in the following oral exercises.

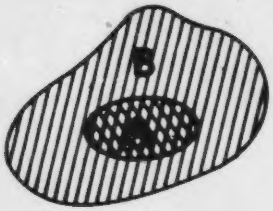


Figure 1a.

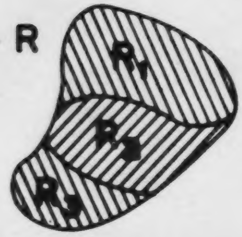


Figure 1b.

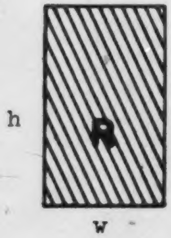


Figure 1c.

Oral Exercises

1. In Figure 1a, if the area of the doubly shaded region B is 2, then what can you say about the entire shaded region A?
2. In Figure 1b, if the areas of regions R_1, R_2, R_3 are respectively 4, 3, and 2, then what is the area of their union R?
3. In Figure 1c, if $h = 5$ and $w = 3$, then what is the area of the rectangle R?
4. What is the smallest area that any region could have?

If you have answered these questions as we believe you have, then you are probably willing to accept as true the following intuitive ideas concerning area.

Property 1. If A and B are two regions with $A \subset B$, then area of $A \leq$ area of B.

Property 2. If a region R is the union of several non-overlapping regions $R_1, R_2, \ldots, R_n$, then area of R = area of R_1 + area of R_2 + ... + area of R_n.

Property 3. The area of a rectangle is the product of the length and the width.

Property 4. The area of any region is greater than or equal to zero.

We will assume these properties to be true from this time on. In order to see where these assumptions lead, we turn to the example in the introduction.

Example 1. Here we are attempting to find some information concerning the area of the region R under the curve

$$f(x) = x^2$$

between $x = 0$ and $x = 1$, the shaded region in Figure 2a.

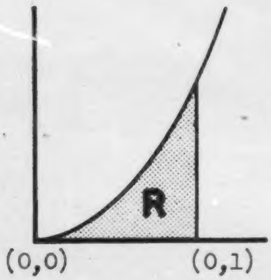


Figure 2a

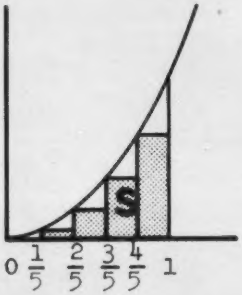


Figure 2b

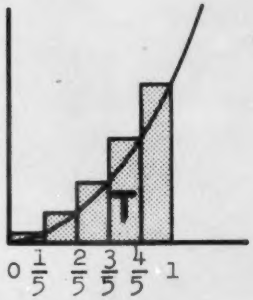


Figure 2c

In Figure 2b a region S has been drawn with $S \subset R$. In Figure 2c another region T has been drawn with $R \subset T$. Since

$$S \subset R \subset T$$

we see from Property 1 that (if R does indeed have an area)

$$\text{area of } S \leq \text{area of } R \leq \text{area of } T. \qquad (1)$$

The region T is seen to be the union of five non-overlapping rectangles, T_1, T_2, T_3, T_4, T_5. Hence, by Property 2,

$$\text{area of } T = \text{area of } T_1 + \text{area of } T_2 + \ldots + \text{area of } T_5. \qquad (2)$$

Each of these rectangles has width equal to $\frac{1}{5}$, and their heights are, respectively,

$$f(\tfrac{1}{5}),\ f(\tfrac{2}{5}),\ f(\tfrac{3}{5}),\ f(\tfrac{4}{5}),\ f(\tfrac{5}{5}),$$

that is,

$$\frac{1}{25},\ \frac{4}{25},\ \frac{9}{25},\ \frac{16}{25},\ \frac{25}{25}.$$

Thus by Property 3, the areas of these rectanges are, respectively,

$$\frac{1}{5}\cdot\frac{1}{25},\ \frac{1}{5}\cdot\frac{4}{25},\ \frac{1}{5}\cdot\frac{9}{25},\ \frac{1}{5}\cdot\frac{16}{25},\ \frac{1}{5}\cdot\frac{25}{25}.$$

Therefore, by (2),

$$\begin{aligned}\text{area of } T &= \frac{1}{5}\cdot\frac{1}{25}+\frac{1}{5}\cdot\frac{4}{25}+\frac{1}{5}\cdot\frac{9}{25}+\frac{1}{5}\cdot\frac{16}{25}+\frac{1}{5}\cdot\frac{25}{25}\\ &= \frac{1}{125}\cdot(1+4+9+16+25). \qquad (3)\end{aligned}$$

This discussion may seem a little drawn out, but our purpose has been to show exactly where the Properties 1, 2, 3 were used.

In the same way S is seen to be the union of five non-overlapping rectangles (we have included the "degenerate" rectangle consisting of the segment from $(0,0)$ to $(\frac{1}{5}, 0)$ which has height equal to zero). The area of S is thus seen to be given by

$$\text{area of } S = \frac{1}{125}(0+1+4+9+16). \qquad (4)$$

Now, (1), (3), and (4) yield

$$\frac{1}{125}(0+1+4+9+16) \le \text{area of } R \le \frac{1}{125}(1+4+9+16+25)$$

or

$$\frac{30}{125} \le \text{area of } R \le \frac{55}{125}$$

or

$$\frac{6}{25} \le \text{area of } R \le \frac{11}{25}.$$

Taking the average of the upper and lower estimates, we find that .34 is an approximation of the area of R with error less than .1, that is,

$$|(\text{area of } R) - .34| \le .1\,.$$

Taking a look at what we have done, we find that accepting the Properties 1-4 has forced us to the conclusion that the area of R must be between $\frac{6}{25}$ and $\frac{11}{25}$. It is reasonable to expect that, if the interval $[0,1]$ were divided into a larger number of pieces, then a better approximation for the area of R would result. This is not only reasonable to expect, but it is very easy to show by means of the following pictures.

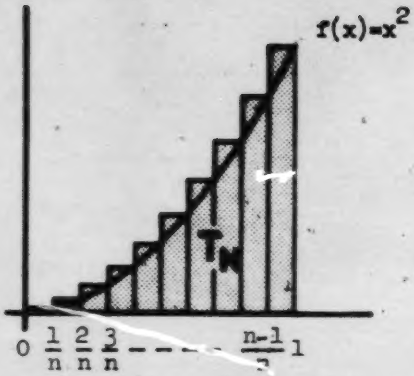


Figure 3a.

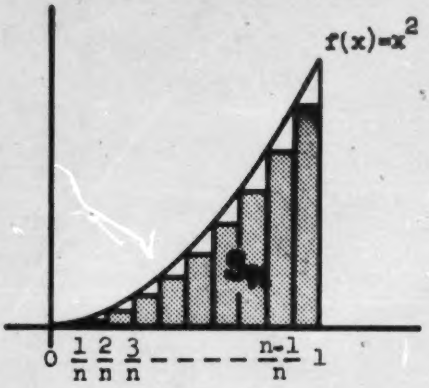


Figure 3b.

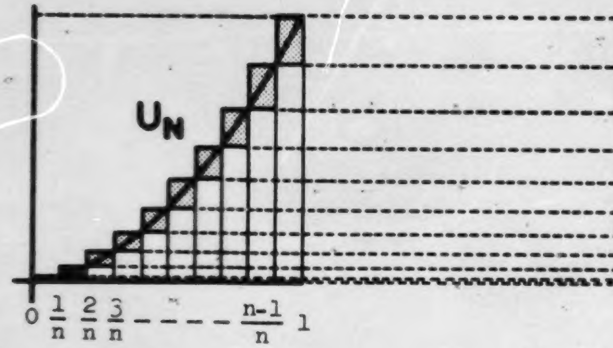


Figure 3c.

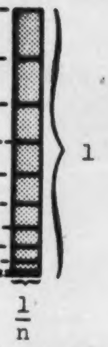


Figure 3d.

In Figure 3a and Figure 3b we have constructed regions T_n and S_n, respectively, containing and contained in R. Hence

$$S_n < R < T_n$$

and

$$\text{area of } S_n < \text{area of } R < \text{area of } T_n.$$

The region U_n shaded in Figure 3c is what is left of T_n after S_n is deleted. Thus

$$\text{area of } U_n = (\text{area of } T_n) - (\text{area of } S_n).$$

Figure 3d shows that U_n has area equal to that of a rectangle with height 1 and width $\frac{1}{n}$, so that

$$\text{area of } U_n = \frac{1}{n}.$$

Since

$$\text{area of } S_n < \text{area of } R < \text{area of } T_n$$

we see now that we have the area of R bracketed between two numbers which differ by $\frac{1}{n}$. It is clear that by choosing n large enough we can get as close an estimate as we like for the value of the area of R. Furthermore, if we can compute the area of T_n, then the area of S_n may be obtained by subtracting $\frac{1}{n}$.

Instead of choosing a particular value of n and evaluating the area of T_n for this n, we should like to find a formula yielding the value of the area of T_n for all n. To this end we observe that

$$\begin{aligned} \text{area of } T_n &= \frac{1}{n} f\left(\frac{1}{n}\right) + \frac{1}{n} f\left(\frac{2}{n}\right) + \frac{1}{n} f\left(\frac{3}{n}\right) + \dots + \frac{1}{n} f\left(\frac{n}{n}\right) \\ &= \frac{1}{n}\left(\frac{1}{n}\right)^2 + \frac{1}{n}\left(\frac{2}{n}\right)^2 + \frac{1}{n}\left(\frac{3}{n}\right)^2 + \dots + \frac{1}{n}\left(\frac{n}{n}\right)^2. \end{aligned}$$

In the ensuing pages, sums similar to the above will occur over and over again. Because such sums have a way of stretching all the way across the page and being most difficult to read, we digress to introduce a notation to simplify the writing of such sums.

6.2 Sigma Notation

Near the end of the preceding section we encountered the formula

$$\text{area of } T_n = \frac{1}{n} f\left(\frac{1}{n}\right) + \frac{1}{n} f\left(\frac{2}{n}\right) + \frac{1}{n} f\left(\frac{3}{n}\right) + \dots + \frac{1}{n} f\left(\frac{n}{n}\right).$$

In the sum on the right, all the terms have the same form, that is, they can all be expressed in the form

$$\frac{1}{n} f\left(\frac{k}{n}\right) \quad \text{for } k = 1, 2, 3, \dots, n.$$

Now we introduce the notation

$$\sum_{k=1}^{n} \frac{1}{n} f\left(\frac{k}{n}\right)$$

to stand for the sum

$$\frac{1}{n} f\left(\frac{1}{n}\right) + \frac{1}{n} f\left(\frac{2}{n}\right) + \frac{1}{n} f\left(\frac{3}{n}\right) + \dots + \frac{1}{n} f\left(\frac{n}{n}\right).$$

Similarly,

$$\sum_{k=4}^{9} k^2 = 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2 = 271.$$

In general, if m and n are integers with $m \leq n$, we define

$$\sum_{k=m}^{n} a_k$$

as follows: we evaluate a_k for each integer k from m to n (inclusive) and add up all the results. Thus

$$\sum_{k=5}^{11} k = 5 + 6 + 7 + 8 + 9 + 10 + 11 = 56$$

and

$$\sum_{k=1}^{9} 3 = 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 = 27.$$

Clearly, this notation has nothing to do with the letter k, that is,

$$\sum_{k=1}^{n} a_k = \sum_{j=1}^{n} a_j.$$

It is only necessary to see how various basic properties of addition look in this notation in order to use the notation properly. The distributive law, for example, becomes:

$$\sum_{k=1}^{n} ca_k = c \sum_{k=1}^{n} a_k.$$

This holds true so long as c does not depend on k. If c depends on n, the formula is still true since n is fixed; it is only k which assumes different values. The addition property of inequalities assumes the form,

$$\text{if for all integers } k \text{ from } 1 \text{ to } n \quad a_k \le b_k,$$

$$\text{then} \quad \sum_{k=1}^{n} a_k \le \sum_{k=1}^{n} b_k.$$

The commutative and associative properties yield

$$\sum_{k=1}^{n} (a_k + b_k) = \left(\sum_{k=1}^{n} a_k\right) + \left(\sum_{k=1}^{n} b_k\right).$$

That is to say, the sum of n number of terms is independent of the way the terms are rearranged or grouped.

One last observation which is useful over and over again in actually evaluating sums is that

$$\sum_{k=1}^{n} (a_k - a_{k-1}) = a_n - a_0.$$

This is easily seen as follows

$$\begin{aligned}\sum_{k=1}^{n} (a_k - a_{k-1}) &= (a_1 - a_0) + (a_2 - a_1) + (a_3 - a_2) + \dots + (a_n - a_{n-1}) \\ &= (-a_0 + a_1) + (-a_1 + a_2) + (-a_2 + a_3) + \dots + (-a_{n-1} + a_n) \\ &= -a_0 + a_n\end{aligned}$$

because all the intermediate terms cancel out.

Now let us evaluate a few sums. For example: $1 + 2 + 3 + \dots + n$. This sum is expressed in our notation as $\sum_{k=1}^{n} k$. Now the summand, k, can be expressed in the form $a_k - a_{k-1}$ with $a_k = \frac{1}{2}(k + \frac{1}{2})^2$ since

$$\begin{aligned}a_k - a_{k-1} &= \tfrac{1}{2}(k + \tfrac{1}{2})^2 - \tfrac{1}{2}(k - \tfrac{1}{2})^2 \\ &= \tfrac{1}{2}(k^2 + k + \tfrac{1}{4}) - \tfrac{1}{2}(k^2 - k + \tfrac{1}{4}) \\ &= k.\end{aligned}$$

Thus

$$\sum_{k=1}^{n} k = \sum_{k=1}^{n} (a_k - a_{k-1})$$

with a_k defined by $a_k = \frac{1}{2}(k + \frac{1}{2})^2$. Thus

$$\begin{aligned}\sum_{k=1}^{n} k = a_n - a_0 &= \tfrac{1}{2}(n + \tfrac{1}{2})^2 - \tfrac{1}{2}(\tfrac{1}{2})^2 \\ &= \tfrac{1}{2}(n^2 + n + \tfrac{1}{8}) - \tfrac{1}{8} \\ &= \frac{n^2 + n}{2} \\ &= \frac{n(n + 1)}{2}.\end{aligned}$$

Similarly,

$$\sum_{k=1}^{n} k^2$$

can be evaluated by first observing that

$$(k + \tfrac{1}{2})^3 - (k - \tfrac{1}{2})^3 = (k^3 + \tfrac{3}{2}k^2 + \tfrac{3}{4}k + \tfrac{1}{8}) - (k^3 - \tfrac{3}{2}k^2 + \tfrac{3}{4}k - \tfrac{1}{8}) = 3k^2 + \tfrac{1}{4}.$$

Now we can evaluate the sum

$$\sum_{k=1}^{n} k^2$$

somewhat indirectly by first evaluating

$$\sum_{k=1}^{n} (3k^2 + \tfrac{1}{4}).$$

To this end we note that

$$3k^2 + \tfrac{1}{4} = a_k - a_{k-1}$$

where $a_k = (k + \frac{1}{2})^3$ so that

$$a_{k-1} = (k - 1 + \tfrac{1}{2})^3 = (k - \tfrac{1}{2})^3.$$

Thus

$$\begin{aligned}\sum_{k=1}^{n} (3k^2 + \tfrac{1}{4}) = \sum_{k=1}^{n} (a_k - a_{k-1}) = a_n - a_0 &= (n + \tfrac{1}{2})^3 = (\tfrac{1}{2})^3 \\ &= n^3 + \tfrac{3}{2}n^2 + \tfrac{3}{4}n + \tfrac{1}{8} - \tfrac{1}{8} = n^3 + \tfrac{3}{2}n^2 + \tfrac{3}{4}n.\end{aligned}$$

On the other hand

$$\sum_{k=1}^{n} (3k^2 + \tfrac{1}{4}) = \sum_{k=1}^{n} 3k^2 + \sum_{k=1}^{n} \tfrac{1}{4} = 3 \sum_{k=1}^{n} k^2 + \tfrac{1}{4}n.$$

Combining the two expressions for

$$\sum_{k=1}^{n} (3k^2 + \tfrac{1}{4})$$

we have

$$3\sum_{k=1}^{n} k^2 + \frac{1}{4}n = n^3 + \frac{3}{2}n^2 + \frac{3}{4}n.$$

Solving for $\sum_{k=1}^{n} k^2$, we have

$$3\sum_{k=1}^{n} k^2 = n^3 + \frac{3}{2}n^2 + \frac{1}{2}n$$

$$\sum_{k=1}^{n} k^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

$$= \frac{2n^3 + 3n^2 + n}{6}$$

$$= \frac{n(n+1)(2n+1)}{6}.$$

The basic idea here was to get some quadratic expression in k (in this case $3k^2 + 1$) expressed in the form $a_k - a_{k-1}$. The rest of the computation was entirely straightforward. As another example try

$$\sum_{k=1}^{n} k^3.$$

Note that

$$(k + \tfrac{1}{2})^4 - (k - \tfrac{1}{2})^4 = k^4 + 2k^3 + \tfrac{3}{2}k^2 + \tfrac{1}{2}k + \tfrac{1}{8} - (k^4 - 2k^3 + \tfrac{3}{2}k^2 - \tfrac{1}{2}k + \tfrac{1}{8}) = 4k^3 + k.$$

Now,

$$\sum_{k=1}^{n} (4k^3 + k) = \sum_{k=1}^{n} [(k + \tfrac{1}{2})^4 - (k - \tfrac{1}{2})^4] = (n + \tfrac{1}{2})^4 - (\tfrac{1}{2})^4.$$

On the other hand

$$\sum_{k=1}^{n} (4k^3 + k) = \sum_{k=1}^{n} 4k^3 + \sum_{k=1}^{n} k = 4\sum_{k=1}^{n} k^3 + \frac{n(n+1)}{2}.$$

Equating the two expressions for

$$\sum_{k=1}^{n} (4k^3 + k)$$

we have

$$4\sum_{k=1}^{n} k^3 + \frac{n(n+1)}{2} = (n + \tfrac{1}{2})^4 - (\tfrac{1}{2})^4$$

$$= [(n + \tfrac{1}{2})^2 + (\tfrac{1}{2})^2][(n + \tfrac{1}{2})^2 - (\tfrac{1}{2})^2]$$

$$= [n^2 + n + \tfrac{1}{2}][n(n+1)].$$

Now,

$$4\sum_{k=1}^{n} k^3 = [n^2 + n + \tfrac{1}{2}][n(n+1)] - \tfrac{1}{2}[n(n+1)]$$

$$= [n^2 + n][n(n+1)] = [n(n+1)]^2.$$

Thus

$$\sum_{k=1}^{n} k^3 = \left(\frac{n(n+1)}{2}\right)^2.$$

Here is a trigonometric example:

$$\sum_{k=1}^{n} \cos 2kA.$$

We recall that

$$\begin{aligned} &\sin(2k+1)A - \sin(2k-1)A \\ &\qquad = \sin 2kA \cos A + \cos 2kA \sin A - (\sin 2kA \cos A - \cos 2kA \sin A) \\ &\qquad = 2 \cos 2kA \sin A. \end{aligned}$$

Therefore

$$\cos 2kA = \frac{1}{2 \sin A}\left(\sin(2k+1)A - \sin(2k-1)A\right).$$

Thus

$$\begin{aligned} \sum_{k=1}^{n} \cos 2kA &= \sum_{k=1}^{n} \frac{1}{2 \sin A}\left(\sin(2k+1)A - \sin(2k-1)A\right) \\ &= \frac{1}{2 \sin A} \sum_{k=1}^{n} \left(\sin(2k+1)A - \sin(2k-1)A\right) \\ &= \frac{1}{2 \sin A} \sum_{k=1}^{n} (a_k - a_{k-1}) \end{aligned}$$

where $a_k = \sin(2k+1)A$. Therefore

$$\sum_{k=1}^{n} \cos 2kA = \frac{1}{2 \sin A}(a_n - a_0) = \frac{1}{2 \sin A}\left(\sin(2n+1)A - \sin A\right).$$

An especially simple example is

$$\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \dots + \frac{1}{n(n+1)}$$

or

$$\sum_{k=1}^{n} \frac{1}{k(k+1)}.$$

Now

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}.$$

Thus

$$\sum_{k=1}^{n} \frac{1}{k(k+1)} = -\sum_{k=1}^{n} \left(\frac{1}{k+1} - \frac{1}{k}\right) = -\sum_{k=1}^{n} (a_k - a_{k-1})$$

where $a_k = \frac{1}{k+1}$. Thus

$$\sum_{k=1}^{n} \frac{1}{k(k+1)} = -(a_n - a_0) = -\left(\frac{1}{n+1} - \frac{1}{1}\right) = \frac{n}{n+1}.$$

We can sum the familiar geometric progression

$$r + r^2 + r^3 + \dots + r^n$$

by the same method. Consider

$$\sum_{k=1}^{n} (r^{k+1} - r^k) = r^{n+1} - r^1.$$

But

$$\sum_{k=1}^{n} (r^{k+1} - r^k) = \sum_{k=1}^{n} r^k(r - 1) = (r - 1) \sum_{k=1}^{n} r^k.$$

Thus

$$(r - 1) \sum_{k=1}^{n} r^k = r^{n+1} - r$$

so that

$$\sum_{k=1}^{n} r^k = \frac{r^{n+1} - r}{r - 1}.$$

You should not get the idea that we will always be successful in finding simple expressions for sums of the form $\sum_{k=1}^{n} a_k$. No amount of work will suffice to produce such formulas for the sums

$$\sum_{k=1}^{n} \frac{1}{k} \quad \text{or} \quad \sum_{k=1}^{n} \frac{1}{k^2}.$$

6.3 Computation of Areas

After the digression of Section 6.2 in order to introduce a convenient notation, we return to the area problem of Section 6.1. We were in the process of finding an estimate for the area of the region R under the parabola $f(x) = x^2$ between $x = 0$ and $x = 1$.

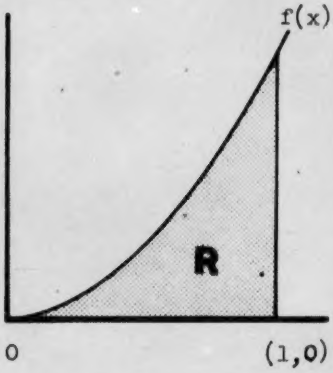


Figure 4a.

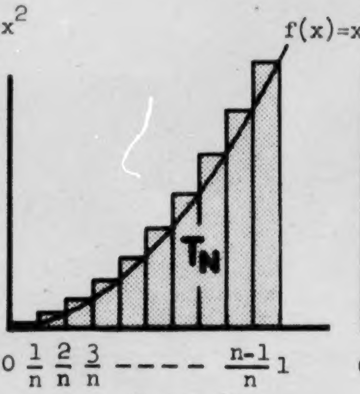


Figure 4b.

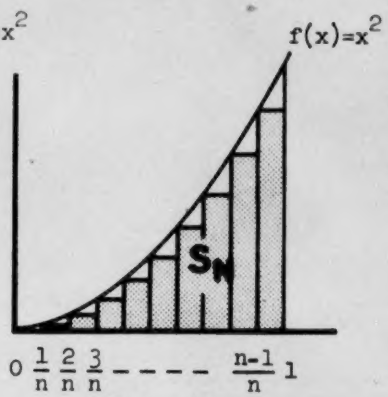


Figure 4c.

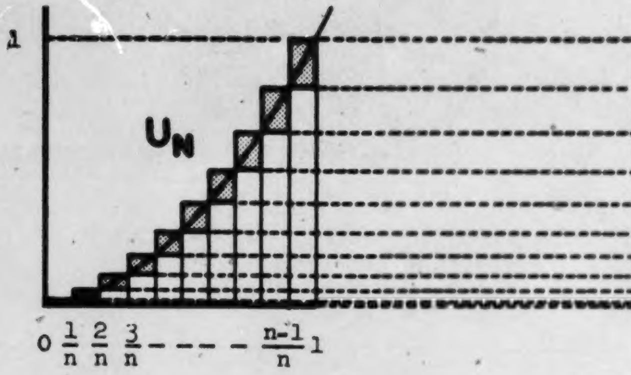

1
U_N
$0 \ \frac{1}{n} \ \frac{2}{n} \ \frac{3}{n} - - - - - \frac{n-1}{n} \ 1$

Figure 4d.

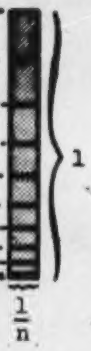


Figure 4e.

The above pictures should remind you that we had shown that

$$\text{area of } S_n < \text{area of } R < \text{area of } T_n,$$

that

$$\text{area of } U_n = \text{area of } T_n - \text{area of } S_n = \frac{1}{n}$$

and that

$$\text{area of } T_n = \frac{1}{n} f\left(\frac{1}{n}\right) + \frac{1}{n} f\left(\frac{2}{n}\right) + \frac{1}{n} f\left(\frac{3}{n}\right) + \dots + \frac{1}{n} f\left(\frac{n}{n}\right).$$

In our new notation this formula may be expressed

$$\text{area of } T_n = \sum_{k=1}^{n} \frac{1}{n} f\left(\frac{k}{n}\right) = \sum_{k=1}^{n} \frac{1}{n} \left(\frac{k}{n}\right)^2.$$

The sum $\sum_{k=1}^{n} \frac{1}{n} \left(\frac{k}{n}\right)^2$ may be evaluated in the following manner:

$$\sum_{k=1}^{n} \frac{1}{n} \left(\frac{k}{n}\right)^2 = \sum_{k=1}^{n} \frac{k^2}{n^3} = \frac{1}{n^3} \sum_{k=1}^{n} k^2.$$

Now the sum $\sum_{k=1}^{n} k^2$ was shown in the preceding section to be equal to

$$\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}.$$

So we find that

$$\text{area of } T_n = \frac{1}{n^3} \left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right) = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}$$

and

$$\text{area of } S_n = (\text{area of } T_n) - \frac{1}{n} = \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2}.$$

Therefore for any positive integer n we must have

$$\frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} < \text{area of } R < \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}.$$

It is clear that we can make the upper and lower estimates for the area of R as close to $\frac{1}{3}$ as we wish by choosing n large enough. It is therefore clear that the area of R must be $\frac{1}{3}$ if it is to satisfy this inequality for all positive integers n.

If we step back and take a look at what we have done, we find that the acceptance of the properties of area at the beginning of this chapter have led us to the conclusion that the area of k can only be $\frac{1}{3}$.

We can apply the above technique to other regions, of course. Here follows a particularly interesting example.

Let us attempt to find the area under the curve

$$f(x) = \cos x$$

between $x = 0$ and $x = a$ where a may be any number between 0 and $\frac{\pi}{2}$.

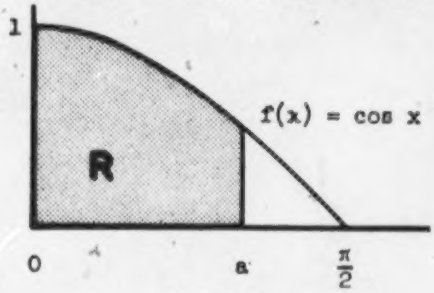

1
f(x) = cos x
R
0
a
π/2

Figure 5.

We divide the interval $[0,a]$ into n subintervals of equal length and construct the "containing" and "contained" rectangles.

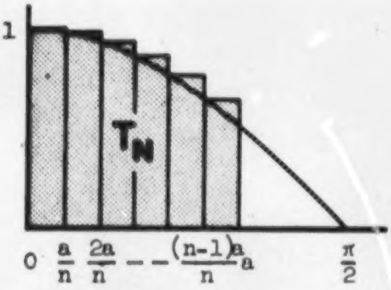

1
T_N
0
a/n
2a/n
(n-1)a/n
a
π/2

Figure 6a.

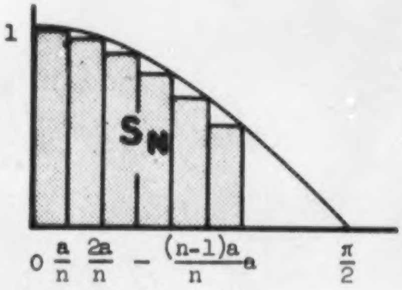

1
S_N
0
a/n
2a/n
(n-1)a/n
a
π/2

Figure 6b.

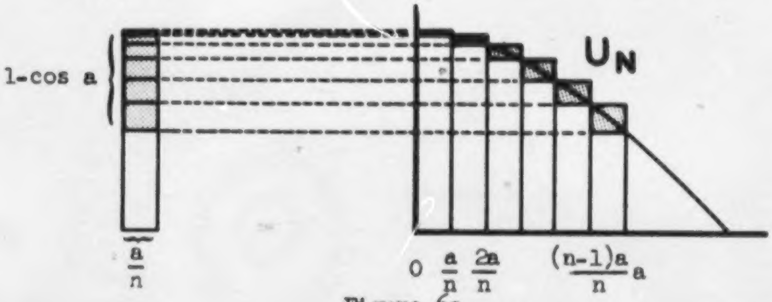

1-cos a
U_N
a/n
0
a/n
2a/n
(n-1)a/n
a

Figure 6c.

As in the preceding example, we have

$$S_n < R < T_n$$

so that

$$\text{area of } S_n < \text{area of } R < \text{area of } T_n$$

as illustrated in Figures 6a and 6b. In Figure 6c we see the set U_n which results when S_n is taken away from T_n. We see that

$$\text{area of } U_n = \frac{a}{n}(1 - \cos a)$$

or

$$(\text{area of } T_n) - (\text{area of } S_n) = \frac{a}{n}(1 - \cos a).$$

This formula shows that the area of S_n and the area of T_n can be made as close together as we wish by taking n sufficiently large. It also shows that once the area of S_n has been computed, the area of T_n is obtained by adding $\frac{a}{n}(1 - \cos a)$. The area of S_n is given by

$$\text{area of } S_n = \sum_{k=1}^{n} \frac{a}{n} f\left(\frac{ka}{n}\right) = \sum_{k=1}^{n} \frac{a}{n} \cos \frac{ka}{n} = \frac{a}{n} \sum_{k=1}^{n} \cos \frac{ka}{n}.$$

Now, in the preceding section we found that

$$\sum_{k=1}^{n} \cos 2kA = \frac{1}{2 \sin A}\left(\sin (2n + 1)A - \sin A\right) = \frac{1}{2 \sin A} \sin (2n + 1)A - \frac{1}{2}.$$

The sum $\sum_{k=1}^{n} \cos \frac{ka}{n}$ has exactly this form with $A = \frac{a}{2n}$. Thus,

$$\sum_{k=1}^{n} \cos \frac{ka}{n} = \frac{1}{2 \sin \frac{a}{2n}} \sin (2n + 1) \frac{a}{2n} - \frac{1}{2} = \frac{1}{2 \sin \frac{a}{2n}} \sin \left(a + \frac{a}{2n}\right) - \frac{1}{2}.$$

Thus

$$\text{area of } S_n = \frac{a}{n} \sum_{k=1}^{n} \cos \frac{ka}{n} = \frac{\frac{a}{2n}}{\sin \frac{a}{2n}} \sin \left(a + \frac{a}{2n}\right) - \frac{a}{2n}.$$

Now we should like to see what can be said about the area of S_n as n becomes very large. We express

$$\frac{\frac{a}{2n}}{\sin \frac{a}{2n}} \quad \text{in the form} \quad \frac{1}{\frac{\sin \frac{a}{2n}}{\frac{a}{2n}}}.$$

We recall that $\lim_{x \to 0} \frac{\sin x}{x} = 1$. Therefore as n gets large without bound, $\frac{\sin \frac{a}{2n}}{\frac{a}{2n}}$ approaches 1. Also, since the $\sin$ function is continuous, we see that $\sin \left(a + \frac{a}{2n}\right)$ approaches $\sin a$ as n gets large without bound. Now we see that as n becomes large without bound

$$\underset{1}{\frac{\frac{a}{2n}}{\sin\frac{a}{2n}}}\cdot\underset{\sin a}{\sin\left(a+\frac{a}{2n}\right)}-\underset{0}{\frac{a}{2n}}.$$

Therefore, as n gets large without bound,

$$\text{area of } S_n \to 1\cdot\sin a - 0 = \sin a.$$

Similarly,

$$\text{area of } T_n = \text{area of } S_n + \frac{a}{n}(1-\cos a)$$

approaches $\sin a$ also. Since

$$\text{area of } S_n < \text{area of } R < \text{area of } T_n,$$

we see that we have the area of R squeezed between two numbers which can be made as close to $\sin a$ as may be desired. If follows that the area of R can only be $\sin a$.

Again we have found that the assumptions about area made at the beginning of this chapter have forced us to the conclusion that the area of the region R in this problem must be $\sin a$.

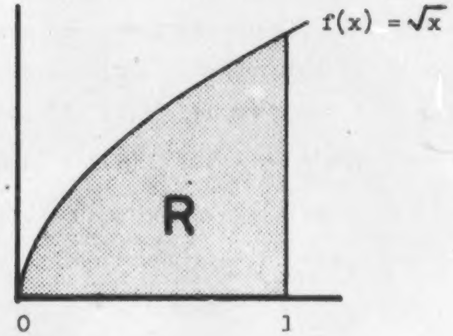
We may not always wish to subdivide our interval into subintervals of equal length. For example, in approximating the area under the curve

$$f(x) = \sqrt{x}$$

between $x = 0$ and $x = 1$, it will be convenient to use such a partition as

$$\{0,\ (\tfrac{1}{5})^2,\ (\tfrac{2}{5})^2,\ (\tfrac{3}{5})^2,\ (\tfrac{4}{5})^2,\ 1\}$$

since the value of $\sqrt{x}$ at these partition points is so easy to compute.



For this partition the upper sum is

$$\frac{1}{25}\sqrt{\frac{1}{25}}+\frac{3}{25}\sqrt{\frac{4}{25}}+\frac{5}{25}\sqrt{\frac{9}{25}}+\frac{7}{25}\sqrt{\frac{10}{25}}+\frac{9}{25}\sqrt{1}=\frac{1+6+15+28+45}{125}=\frac{95}{125}=\frac{19}{25}$$

and the lower sum is $\frac{50}{125} = \frac{10}{25}$. The area under this curve satisfies

$$\frac{10}{25} < \text{area of } R < \frac{19}{25}.$$

We can do better using the partition

$$\{0, (\tfrac{1}{n})^2, (\tfrac{2}{n})^2, (\tfrac{3}{n})^2, \ldots, (\tfrac{n-1}{n})^2, 1\}.$$

The length of the k^{th} subinterval is

$$(\tfrac{k}{n})^2 - (\tfrac{k-1}{n})^2 = \frac{2k-1}{n^2}$$

so that the upper sum corresponding to this partition is

$$\sum_{k=1}^{n} \frac{2k-1}{n^2} f\left(\frac{k^2}{n^2}\right) = \sum_{k=1}^{n} \frac{2k-1}{n^2} \cdot \sqrt{\frac{k^2}{n^2}} = \sum_{k=1}^{n} \frac{2k^2-k}{n^3} = \frac{2}{n^3} \sum_{k=1}^{n} k^2 - \frac{1}{n^3} \sum_{k=1}^{n} k.$$

In the preceding section we found that

$$\sum_{k=1}^{n} k^2 = \frac{n(n-1)(2n+1)}{6} \quad \text{and} \quad \sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

Therefore the upper sum is seen to be

$$\frac{1}{n^3}\left[2\,\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2}\right] = \frac{n(n+1)(4n-1)}{6n^3} = \frac{4n^2+3n-1}{6n^2}.$$

By a similar computation the lower sum is seen to be

$$\sum_{k=1}^{n} \frac{2k-1}{n^2} f\left(\frac{(k-1)^2}{n^2}\right) = \sum_{k=1}^{n} \frac{2k-1}{n^2} \sqrt{\frac{(k-1)^2}{n^2}} = \frac{4n^2-3n-1}{6n^2}.$$

Thus for this region we find

$$\frac{4n^2-3n-1}{6n^2} < \text{area of } R < \frac{4n^2+3n-1}{6n^2}$$

or

$$\frac{2}{3} - \frac{1}{2n} - \frac{1}{6n^2} < \text{area of } R < \frac{2}{3} + \frac{1}{2n} - \frac{1}{6n^2}.$$

We can see that by taking n sufficiently large the upper and lower sums can both be made as close as we wish to $\frac{2}{3}$. The area of R must therefore be $\frac{2}{3}$.

Let us now review and collect our information relating to the problem of finding or approximating the area under a curve. Suppose we have a function f continuous on the interval $[a,b]$, and suppose that $f(x) \geq 0$ for $a \leq x \leq b$. To obtain upper and lower estimates for the area under the curve, we first partition the interval $[a,b]$ by means of numbers (not necessarily equally spaced) $x_0, x_1, x_2, \ldots, x_n$ satisfying

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$

In each of the subintervals $[x_{k-1}, x_k]$, we find the maximum value M_k and minimum value m_k assumed by $f(x)$, that is,

$$M_k = \max f(x) \quad \text{for} \quad x_{k-1} \le x \le x_k$$

$$m_k = \min f(x) \quad \text{for} \quad x_{k-1} \le x \le x_k.$$

Now the sum

$$\sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

is seen to be the sum of areas of non-overlapping rectangles contained in the region under consideration.

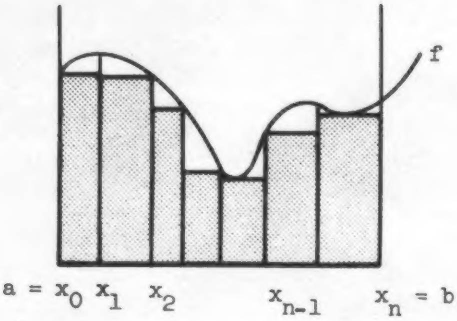


Figure 7a.

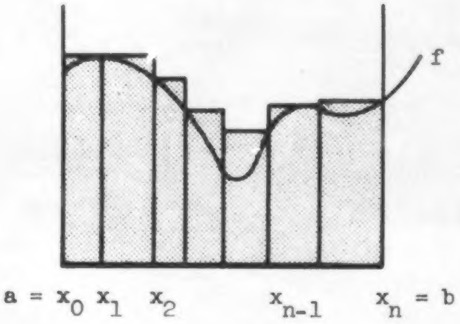


Figure 7b.

The sum

$$\sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

is the sum of the areas of non-overlapping rectangles whose union contains the region under consideration. The values of these sums are determined by the function f and th partition Δ where

$$\Delta = \{x_0, x_1, x_2, \ldots, x_n\}.$$

We will use the notation $\overline{S}_\Delta$ and $\underline{S}_\Delta$ to denote these sums. Thus

$$\overline{S}_\Delta = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

and

$$\underline{S}_\Delta = \sum_{k=1}^{n} m_k(x_k - x_{k-1}).$$

It is true that $\overline{S}_\Delta(f)$ and $\underline{S}_\Delta(f)$ might be better notations as they indicate the dependence on f. In our discussions, however, the intended function f will be clear, and we will stick to the simpler notation. $\overline{S}_\Delta$ is called the upper sum associated with the partition Δ, and $\underline{S}_\Delta$ the lower sum associated with the partition Δ.

Our methods will not always succeed in producing an exact answer to the area problem. For example, in the problem of finding the area under

$$f(x) = \frac{1}{x}$$

between $x = 1$ and $x = 2$, no such answer is available to you at this point. (This matter will be cleared up in the next chapter.) We can still, in such cases, find close approximations to the area.

Suppose that we are given a function f which is monotonely increasing on the interval $[a,b]$. And suppose that we are required to approximate the area under this curve with an error no more than .1.

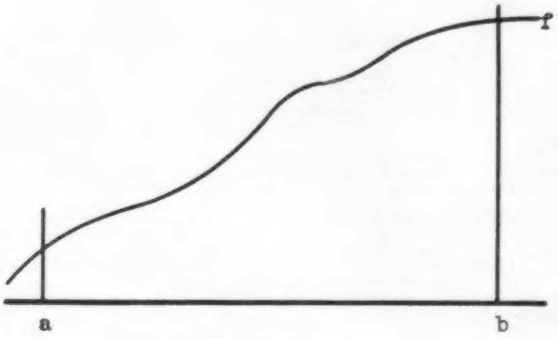


Figure 9.

Our method will be to try to find a partition Δ for which $\overline{S}_\Delta$ and $\underline{S}_\Delta$ differ by no more than .1. If we try to find such a Δ by hit or miss methods, we might be a long time finding it. We should like to have some way of telling in advance whether a partition would do the trick. This turns out to be not at all difficult when the function f is monotone.

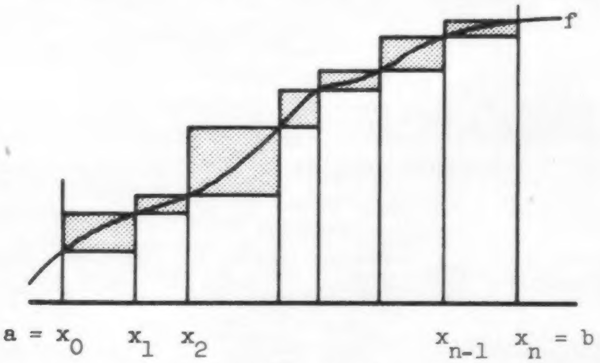


Figure 10.

Figure 10 shows the function of Figure 9 with a partition $\Delta = \{x_0, x_1, \dots, x_n\}$. The shaded region represents the difference

$$\overline{S}_\Delta - \underline{S}_\Delta = \sum_{k=1}^{n} (M_k - m_k)(x_k - x_{k-1}),$$

each of the shaded rectangles being equal in area to a term of this sum. In the picture Figure 10 we see that the subinterval $[x_2, x_3]$ is the widest of all the subintervals formed by this partition. Therefore each of the rectangles of Figure 10 may be "slid over" horizontally (see Figure 11) to fit without overlapping inside the rectangle bounded by the vertical lines $x = x_2$ and $x = x_3$ and the horizontal lines $y = f(a)$ and $y = f(b)$.

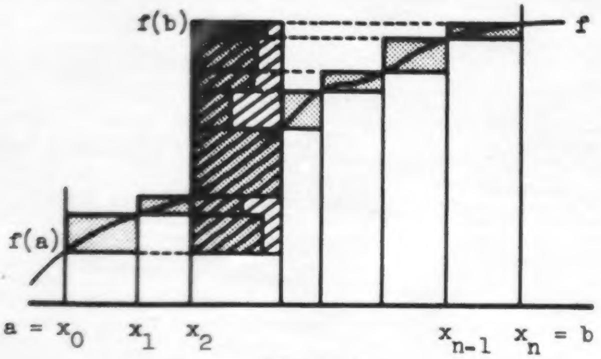


Figure 11.

Thus for the function and the partition of these pictures, we have

$$\overline{S}_\Delta - \underline{S}_\Delta < [f(b) - f(a)](x_3 - x_2).$$

A simple calculation shows that a similar result holds for any monotonely increasing function f and any partition. Let us first introduce the notation $||\Delta||$, called the "norm of Δ," to denote the length of the longest subinterval formed by Δ. That is,

$$||\Delta|| = \text{the maximum of } \{x_1 - x_0, x_2 - x_1, \dots, x_n - x_{n-1}\}.$$

Now we see that

$$\begin{aligned}\overline{S}_\Delta - \underline{S}_\Delta &= \sum_{k=1}^{n} (M_k - m_k)(x_k - x_{k-1}) \\ &\leq \sum_{k=1}^{n} (M_k - m_k)\ ||\Delta|| \\ &= ||\Delta|| \sum_{k=1}^{n} (M_k - m_k).\end{aligned}$$

Next we are ready to use the montoneity of f. Since f is increasing, we see that the maximum functional value on any interval occurs at the right bend endpoint and the minimum at the left. Thus,

$$M_k = f(x_k), \quad m_k = f(x_{k-1}).$$

Now

$$\overline{S}_\Delta - \underline{S}_\Delta \leq ||\Delta|| \sum_{k=1}^{n} (M_k - m_k) = ||\Delta|| \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})].$$

This last sum $\sum_{k=1}^{n} [f(x_k) - f(x_{k-1})]$ is seen to be one of those delightful sums in which almost everything cancels out. Thus

$$\sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] = f(x_n) - f(x_0) = f(b) - f(a).$$

We at last we see that

$$\overline{S}_\Delta - \underline{S}_\Delta \leq ||\Delta|| \, [f(b) - f(a)].$$

(This was illustrated in Figure 11 where $||\Delta||$ was $x_3 - x_2$.) The argument goes the same way for monotonely decreasing function where we would obtain the result

$$\overline{S}_\Delta - \underline{S}_\Delta \leq ||\Delta|| \, [f(b) - f(a)].$$

The two results can be collected in the single statement.

If f is monotone on [a,b], then

$$\overline{S}_\Delta - \underline{S}_\Delta \leq ||\Delta|| \, \left|f(b) - f(a)\right|.$$

The usefulness of this finding is exhibited by such problems as the following.

Problem. Find upper and lower estimates differing by less than .1 for the area under the curve $f(x) = \frac{1}{x}$ over the interval [1,2].

Solution. First, how to choose Δ? Well, we see that $f(a) = \frac{1}{1} = 1$ and $f(b) = \frac{1}{2}$ so that $|f(b) - f(a)| = \frac{1}{2}$. Thus for any partition Δ whatsoever

$$\overline{S}_\Delta - \underline{S}_\Delta \leq ||\Delta|| \, [f(b) - f(a)] = \frac{1}{2} \, ||\Delta||.$$

In order to guarantee that $\overline{S}_\Delta - \underline{S}_\Delta \leq .1$ we need only be sure that Δ is so chosen that $||\Delta|| \leq \frac{1}{5}$. There are many such partitions, of course, the most natural to select being $\Delta = \{1, \frac{6}{5}, \frac{7}{5}, \frac{8}{5}, \frac{9}{5}, 2\}$. The upper and lower sums for this partition are $\overline{S}_\Delta = \frac{1627}{2520}$ and $\underline{S}_\Delta = \frac{1879}{2520}$, and

$$\overline{S}_\Delta - \underline{S}_\Delta = .1 \, .$$

A result similar to

$$\overline{S}_\Delta - \underline{S}_\Delta \le ||\Delta|| \left| f(b) - f(a) \right|$$

is easily obtained for "piecewise monotone" functions. A function f is "piecewise monotone" on [a,b] if there exists a partition of [a,b] such that f is monotone on each of the subintervals of the partition.

Now suppose that f is piecewise monotone on [a,b] and suppose that p is the number of "pieces" into which [a,b] may be partitioned so that f is monotone on the pieces. Let the maximum and minimum values of $f(x)$ on the interval [a,b] be denoted by M and m, respectively. Then for any partition Δ of [a,b] we have

$$\overline{S}_\Delta - \underline{S}_\Delta \le p||\Delta|| \, (M - m).$$

This statement becomes obvious on considering the following pictures.

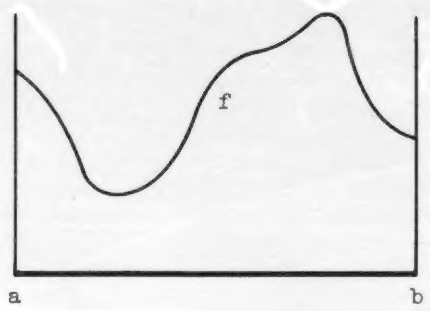


Figure 12a.

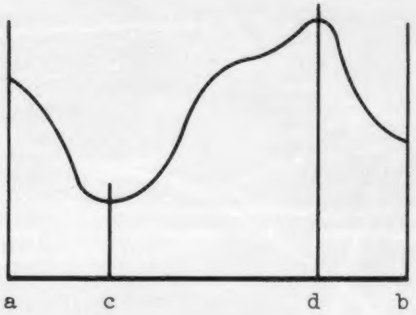


Figure 12b.

Figure 12a shows a function f; Figure 12b shows how the interval [a,b] may be partitioned into three pieces on each of which f is monotone.

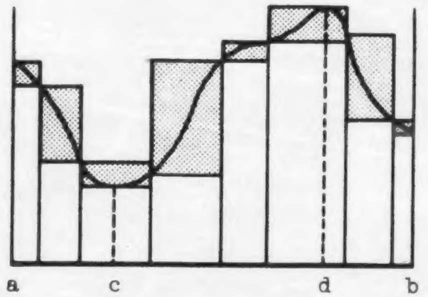


Figure 13a.

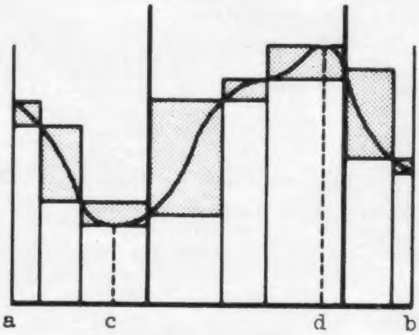


Figure 13b.

In Figure 13a we have added a partition Δ and shaded the area corresponding to $\overline{S}_{\Delta} - \underline{S}_{\Delta}$. In Figure 13b we see how to split the rectangles of Figure 13a into p groups (that is, 3 groups) so that it is obvious that the total area of the rectangles in each group is less than $(M - m) \, ||\Delta||$. The reason that this is so obvious is that for each group the projections of the rectangles on the y-axis do not overlap! Thus, since there are p such groups in all,

$$\overline{S}_{\Delta} - \underline{S}_{\Delta} \leq p||\Delta|| \, (M - m).$$

In speaking of continuous functions, a typical example of which is illustrated below, we are likely to be wary of trusting our intuition.

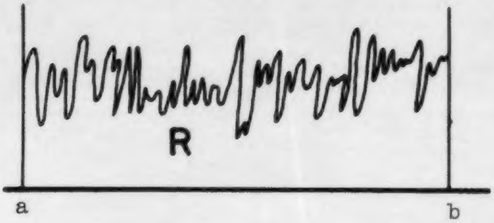


Figure 14.

For example, it seems intuitively reasonable that for any two partitions Δ_1 and Δ_2 of $[a,b]$ we ought to be able to conclude that

$$\underline{S}_{\Delta_2} \leq \overline{S}_{\Delta_1}$$

since

$$\underline{S}_{\Delta_2} < \text{area of } R \quad \text{and} \quad \text{area of } R < \overline{S}_{\Delta_1}.$$

However, can we be sure that for such complicated functions the region R actually has an area? It can be shown independently without resorting to the use of area that

$$\underline{S}_{\Delta_2} \leq \overline{S}_{\Delta_1}.$$

Suppose that f is a function on $[a,b]$ and that Δ is a partition of this interval. Let us see what happens to the upper and lower sums when one additional point x' is adjoined to this partition between x_{j-1} and x_j, thus forming a new partition Δ'.

Now we see that all the terms of

$$\overline{S}_{\Delta} = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

except one also appear in the sum $\overline{S}_{\Delta}'$.

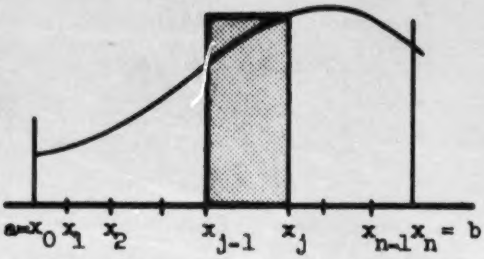


Figure 15a.

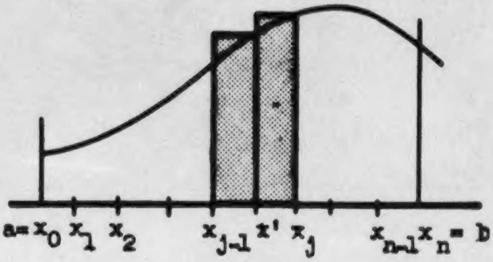


Figure 15b.

One term of $\overline{S}_\Delta$

$$M_j(x_j - x_{j-1})$$

is replaced by

$$M_j'(x' - x_{j-1}) + M_j''(x_j - x')$$

where

$$M_j' = \max f(x) \quad \text{for} \quad x_{j-1} \leq x \leq x'$$

and

$$M_j'' = \max f(x) \quad \text{for} \quad x' \leq x \leq x_j.$$

(See Figs. 15a and 15b.)

It is clear that the maximum value of a function on a subinterval cannot be larger than the maximum on the whole interval. Thus

$$M_j' \leq M_j \quad \text{and} \quad M_j'' \leq M_j$$

so that

$$M_j'(x' - x_{j-1}) + M_j''(x_j - x') \leq M_j(x' - x_{j-1}) + M_j(x_j - x') = M_j(x_j - x_{j-1}).$$

This inequality shows us that the sum of the terms replacing $M_j(x_j - x_{j-1})$ in obtaining $\overline{S}_\Delta'$ from $\overline{S}_\Delta$ cannot be larger than the term they replaced. Therefore

$$\overline{S}_\Delta' \leq \overline{S}_\Delta.$$

In just the same way it can be shown that $m_j' \geq m_j$ and $m_j'' \geq m_j$ so that

$$\underline{S}_\Delta' \geq \underline{S}_\Delta.$$

Now we are ready to see why, for any two partitions Δ_1 and Δ_2 of $[a,b]$, we must have

$$\underline{S}_{\Delta_1} \leq \overline{S}_{\Delta_2}.$$

For, let Δ_3 be the partition consisting of all the partition points of both Δ_1 and Δ_2. Now Δ_3 can be obtained from Δ_2 by starting with Δ_2 and successively adjoining one point of Δ_1 at a time. At no stage of this process is the upper sum ever increased. Thus the final upper sum in this process $\overline{S}_{\Delta_3}$ cannot be greater than the initial one $\overline{S}_{\Delta_2}$. That is,

$$\overline{S}_{\Delta_3} \leq \overline{S}_{\Delta_2}.$$

In the same way we see that

$$\underline{S}_{\Delta_1} \leq \underline{S}_{\Delta_3}.$$

It is, of course, obvious that

$$\underline{S}_{\Delta_3} \leq \overline{S}_{\Delta_3}$$

since the terms in the sum $\underline{S}_{\Delta_3}$ are term by term less than or equal to those in $\overline{S}_{\Delta_3}$. At last we have

$$\underline{S}_{\Delta_1} \leq \underline{S}_{\Delta_3} \leq \overline{S}_{\Delta_3} \leq \overline{S}_{\Delta_2}.$$

The result we have just obtained can be formulated as follows. Suppose that we have a function f continuous on [a,b]. Suppose we let $\mathcal{L}$ represent the set of all lower sums and $\mathcal{U}$ represent the set of all upper sums. Then every member of $\mathcal{L}$ is less than or equal to every member of $\mathcal{U}$.

When this result is formulated in this way, we see that we can assert, from Property 8 of the real numbers, that there is at least one number S separating the two sets $\mathcal{L}$ and $\mathcal{U}$. That is, for any members $\underline{S}_{\Delta_1}$ of $\mathcal{L}$ and $\overline{S}_{\Delta_2}$ of $\mathcal{U}$ we will have

$$\underline{S}_{\Delta_1} \leq S \leq \overline{S}_{\Delta_2}.$$

If we were able to show that there is just one such number S which separates $\mathcal{L}$ and $\mathcal{U}$, then we would have just what we want. That is to say that under the assumptions at the beginning of this chapter, there is just one possible value which the area under f over the interval [a,b] could possibly have. Referring again to Chapter 2, we see that in order to demonstrate the uniqueness of the number S separating $\mathcal{L}$ and $\mathcal{U}$, we have only to show that for every positive number ϵ there exist members $\underline{S}_{\Delta_1}$ of $\mathcal{L}$ and $\overline{S}_{\Delta_2}$ of $\mathcal{U}$ such that

$$\overline{S}_{\Delta_2} - \underline{S}_{\Delta_1} < \epsilon.$$

It is interesting to note that, in the case that f is monotone, a considerably stronger result has already been demonstrated, to wit: if f is monotone in [a,b], then for any partition of [a,b] we have

$$\overline{S}_{\Delta} - \underline{S}_{\Delta} < ||\Delta|| \left|f(b) - f(a)\right|.$$

Since $|f(b) - f(a)|$ is fixed, we see that by making $||\Delta||$ sufficiently small, we can force $\overline{S}_\Delta - \underline{S}_\Delta$ to be as small as we wish. A similar result holds for piecewise monotone functions in view of the inequality

$$\overline{S}_\Delta - \underline{S}_\Delta \leq p||\Delta|| \, \Big|f(b) - f(a)\Big|.$$

Here p is the number of subintervals into which $[a,b]$ must be partitioned in order that f should be monotone on the subintervals. Again $|f(b) - f(a)|$ and p are fixed, so that by choosing $||\Delta||$ sufficiently small, we may make $\overline{S}_\Delta - \underline{S}_\Delta$ as small as we like.

The situation is slightly different if we know only that the function f is continuous. In the above discussion we found that for every function which is monotone or piecewise monotone on an interval $[a,b]$ we can find a number k so that

$$\overline{S}_\Delta - \underline{S}_\Delta < k||\Delta||.$$

There do, however, exist continuous functions for which no such k can be found. Nevertheless, it is true that for any function continuous on $[a,b]$ we can make $\overline{S}_\Delta - \underline{S}_\Delta$ as small as we like by choosing $||\Delta||$ sufficiently small. That is,

Theorem 1. If f is continuous on $[a,b]$, then for every $\epsilon > 0$ there is a $\delta > 0$ such that $\overline{S}_\Delta - \underline{S}_\Delta < \epsilon$ whenever $||\Delta|| < \delta$.

We will not prove this theorem in this text, for there are unpleasant technical difficulties in our way. The student should recognize that we have proved the theorem for continuous functions which are piecewise monotone. We shall, however, assume the theorem to be true for all continuous functions. From this assumption we see that it follows that for every continuous function there is a unique number S which separates the sets $\mathcal{L}$ and $\mathcal{U}$. Thus we make the following definition.

Definition 1. If f is continuous on $[a,b]$, we define

$$\int_a^b f(x)\,dx$$

(called the integral of f over the interval $[a,b]$) to be the unique number S such that

$$\underline{S}_{\Delta_1} \leq S \leq \overline{S}_{\Delta_2}$$

whenever Δ_1 and Δ_2 are partitions of $[a,b]$.

We re-emphasize that in order for this definition to make sense, it is necessary to have Theorem 1 at our disposal to ensure the uniqueness of the number S. We remind the student that this theorem has not been proved in this book under the hypothesis that f is continuous. But on the other hand, we further re-emphasize that the theorem has been proved under the additional hypothesis that f is piecewise monotone. Now, all the functions that we will encounter in this course (except for a small number of horrible examples) will be piecewise monotone. Thus, for the functions we will actually encounter, the theorem on which this definition rests has been proved. Nevertheless, we will bow to tradition and state the definition for continuous functions.

A remark on the notation involved in

$$\int_a^b f(x)\,dx$$

is in order. It should be clear that the value of this integral is synonymous with the area under the function f over the interval $[a,b]$ and depends only on the function f and the interval $[a,b]$. In particular x has nothing to do with it whatsoever. Thus,

$$\int_a^b f(x)\,dx \qquad \int_a^b f(y)\,dy \qquad \int_a^b f(t)\,dt$$

all have exactly the same meaning. For that reason some authors instead write

$$\int_a^b f.$$

We will, however, adhere to the more standard notation

$$\int_a^b f(x)\,dx$$

which has, in fact, a number of advantages which will become apparent shortly. One of these advantages is that this notation permits us to write

"$\int_a^b x^2\,dx$"

instead of

"$\int_a^b f$

where f is defined by

$f(x) = x^2$."

Letting $\epsilon < 0$ and choosing $\delta < 0$ such that $\overline{S}_\Delta - \underline{S}_\Delta < \epsilon$ whenever $||\Delta|| < \delta$ and recalling that $\underline{S}_\Delta < \int_a^b f(x)\,dx < \overline{S}_\Delta$, we see that both the upper sum and the lower sum will differ from the integral by less than ϵ. Thus, if the norm of the partition is sufficiently small, the upper and lower sums will be close approximations of the integral.

6.6 Riemann Sums

Again let f be a function continuous on $[a,b]$, and let

$$\Delta = \{x_0, x_1, \ldots, x_n\}$$

be a partition of $[a,b]$. In the k^{th} subinterval of this partition choose a number ζ_k. Now we have

$$a = x_0 < x_1 < x_2 < x_3 < \ldots < x_{n-1} < x_n = b$$

and

$$x_0 \le \zeta_1 \le x_1 \le \zeta_2 \le x_2 \le \zeta_3 \le x_3 \le \cdots \le x_{n-1} \le \zeta_n \le x_n.$$

Next consider the sum

$$\sum_{k=1}^{n} f(\zeta_k)(x_k - x_{k-1}).$$

This sum is equal to the area of the union of rectangles depicted below.

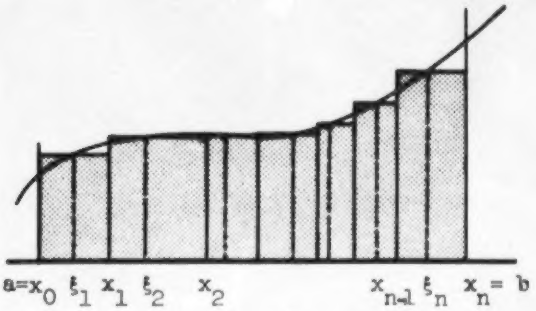


Figure 1.

Such a sum is called a "Riemann sum." Although it is clear that the value of the sum depends on the choice of the numbers $\zeta_1, \zeta_2, \ldots, \zeta_n$, we will nevertheless use the symbol S_Δ to denote such a sum.

For each k we have

$$m_k \le f(\zeta_k) \le M_k$$

so that

$$\underline{S}_\Delta \le \sum_{k=1}^{n} f(\zeta_k)(x_k - x_{k-1}) \le \overline{S}_\Delta.$$

It now follows that

Theorem 2. for every $\epsilon > 0$ there exists a $\delta > 0$ so that, if $||\Delta|| < \delta$, then

$$\left|\sum_{k=1}^{n} f(\zeta_k)(x_k - x_{k-1}) - \int_a^b f(x)\,dx\right| < \epsilon,$$

or

$$|\overline{S}_\Delta - \int_a^b f(x)\, dx| < \epsilon$$

(regardless of how ζ_k may be chosen in the k^{th} subinterval of Δ).

This is made clear by recalling that for every $\epsilon > 0$ we can find $\delta > 0$ so that if $||\Delta|| < \delta$ then $\overline{S}_\Delta - \underline{S}_\Delta < \epsilon$. Since $\int_a^b f(x)\, dx$ and S_Δ both lie between $\underline{S}_\Delta$ and $\overline{S}_\Delta$, it is clear that S_Δ and $\int_a^b f(x)\, dx$ must differ by less than ϵ.

We summarize the state of affairs described in Theorem 2 by saying that

$$\lim_{||\Delta|| \to 0} S_\Delta = \int_a^b f(x)\, dx.$$

We see that we have a different kind of limit here, but one which bears a strong similarity to that encountered in $\lim_{x \to a} f(x)$.

Interpreting the integral as the area under the curve, we can see from Figure 2 that the relation

$$\int_a^c f(x)\, dx = \int_a^b f(x)\, dx + \int_a^c f(x)\, dx$$

ought to hold true when $a < b < c$. This follows from one of our basic assumptions about area at the beginning of this chapter that

$$\text{area of } (R_1 \cup R_2) = \text{area of } R_1 + \text{area of } R_2.$$

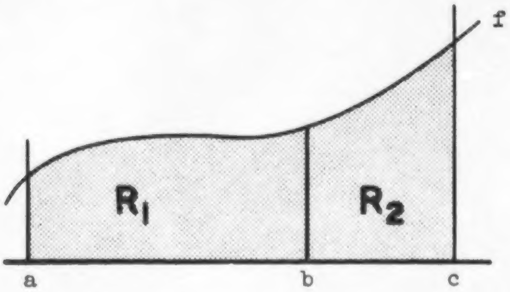


Figure 2.

We should like to be able to establish this relation directly from our analytical definition of the integral more or less as a check on whether this definition truly corresponds to our intuitive concept of area. It turns out that this is quite easy to do, and the method employed is, in fact, rather interesting.

We assume that f is continuous and non-negative on [a,c]. We knew that there is only one number K satisfying

$$\underline{S}_\Delta < K < \overline{S}_\Delta$$

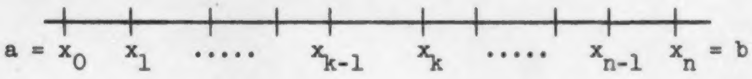
for every partition Δ of $[a,c]$, and this number is $\int_a^c f(x)\,dx$. Consequently if we can establish that

$$\underline{S}_\Delta \le \int_a^b f(x)\,dx + \int_b^c f(x)\,dx \le \overline{S}_\Delta$$

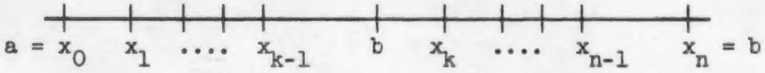
for every partition Δ of $[a,c]$, then it must follow that

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx.$$

It only remains to establish this inequality. Let Δ be any partition of $[a,c]$.



Let Δ_1 be the partition obtained by adjoining the additional partition point b



which falls, say, between x_{k-1} and x_k. Now we know that

$$\underline{S}_\Delta \le \underline{S}_{\Delta_1} \quad \text{and} \quad \overline{S}_{\Delta_1} \le \overline{S}_\Delta.$$

The partitions

$$S_2 = \{x_0, x_1, \ldots, x_{k-1}, b\}$$
$$S_3 = \{b, x_k, \ldots, x_n\}$$

are partitions of $[a,b]$ and $[b,c]$, respectively. The terms in the sum $\underline{S}_{\Delta_1}$ consist of just those terms in the two sums $\underline{S}_{\Delta_2}$ and $\underline{S}_{\Delta_3}$ so that

$$\underline{S}_{\Delta_1} = \underline{S}_{\Delta_2} + \underline{S}_{\Delta_3}.$$

This is illustrated in the following figure.

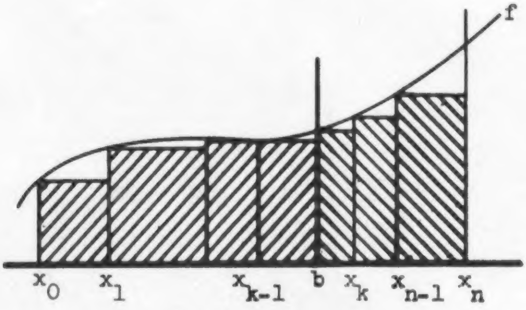


Figure 3.

The entire shaded area represents $\underline{S}_{\Delta_1}$, the regions representing $\underline{S}_{\Delta_2}$ and $\underline{S}_{\Delta_3}$ pairs distinguished in the manner of shading. Similarly for the upper sums

$$\overline{S}_{\Delta_1} = \overline{S}_{\Delta_2} = \overline{S}_{\Delta_3}.$$

Now, of course,

$$\underline{S}_{\Delta_2} \leq \int_a^b f(x)\, dx \leq \overline{S}_{\Delta_2}$$

$$\underline{S}_{\Delta_3} \leq \int_b^c f(x)\, dx \leq \overline{S}_{\Delta_3}.$$

Adding these inequalities in columns yields

$$\underline{S}_{\Delta_2} + \underline{S}_{\Delta_3} \leq \int_a^b f(x)\, dx + \int_b^c f(x)\, dx \leq \overline{S}_{\Delta_2} + \overline{S}_{\Delta_3}.$$

Collecting all we have said yields the string of inequalities

$$\underline{S}_{\Delta} \leq \underline{S}_{\Delta_1} = \underline{S}_{\Delta_2} + \underline{S}_{\Delta_3} \leq \int_a^b f(x)\, dx + \int_b^c f(x)\, dx \leq \overline{S}_{\Delta_2} + \overline{S}_{\Delta_3} = \overline{S}_{\Delta_1} = \overline{S}_{\Delta}.$$

Thus, we have succeeded in showing that

$$\underline{S}_{\Delta} \leq \int_a^b f(x)\, dx + \int_b^c f(x)\, dx \leq \overline{S}_{\Delta}$$

for all partitions Δ of $[a,c]$, and as we have already observed

$$\int_a^b f(x)\, dx + \int_b^c f(x)\, dx = \int_a^c f(x)\, dx$$

in consequence.

STUDENT TEXT

Chapter 8

THE LOGARITHMIC AND EXPONENTIAL FUNCTIONS

In the last chapter we found antiderivatives (indefinite integrals) for x^n for all integers n with the single exception of -1. The formula obtained was

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C \qquad n \neq -1.$$

This formula not only fails to hold when $n = -1$, it doesn't even make sense!

Does the function f defined by

$$f(x) = \frac{1}{x}$$

have an antiderivative? This question is most easily answered. Consider the function L defined by

$$L(x) = \int_1^x \frac{1}{t}\, dt \qquad x > 0.$$

Since $x > 0$, the integral is continuous in the interval between 1 and x. Therefore the integral exists, and the Fundamental Theorem of Calculus applies to yield

$$L'(x) = \frac{1}{x} \qquad x > 0.$$

The question of the existence of an antiderivative of $\frac{1}{x}$ has been answered, but the answer turns out not to be a power of x; the behavior of the function is not obvious. We will devote some time to the study of this function. First we will look at the graph of $y = \frac{1}{x}$, $x > 0$.

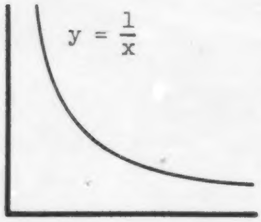


Figure 1a.

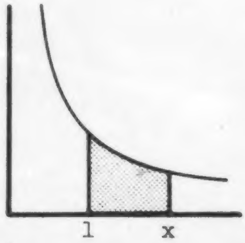


Figure 1b.

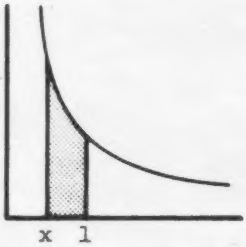


Figure 1c.

The value of $L(x) = \int_1^x \frac{1}{t}\, dt$ is, for $x > 1$, the area under the graph of $y = \frac{1}{x}$ shaded in Figure 1b; while for $0 < x < 1$, $L(x)$ is the negative of the area under this curve between x and 1 as shaded in Figure 1c. It is obvious that $L(1) = \int_1^1 \frac{1}{t}\,dt = 0$, and since the integrand is positive, L is an increasing function.

The most important property of the function L, from which all the subsequent properties are derived, is that

$$L(ab) = L(a) + L(b) \qquad a,b > 0.$$

This can be shown in several ways, for example: Computing $D_x L(ax)$ by means of the chain rule, we have

$$D_x L(ax) = L'(ax) \cdot D_x(ax) = \frac{1}{ax} \cdot a = \frac{1}{x}.$$

Since it is also true that $D_x L(x) = \frac{1}{x}$, we know that $L(ax)$ and $L(x)$ must differ by a constant, thus

$$L(ax) = L(x) + C.$$

We evaluate the constant by setting $x = 1$. Then $L(a) = L(1) + C$ or $L(a) = C$. Therefore

$$L(ax) = L(x) + L(a).$$

This result is sufficiently interesting that we give an additional proof of it.

$$L(ab) = \int_1^{ab} \frac{1}{t}\,dt = \int_1^{a} \frac{1}{t}\,dt + \int_a^{ab} \frac{1}{t}\,dt.$$

Now the integral $\int_1^a \frac{1}{t}\,dt$ is, of course, equal to $L(a)$; to the second integral we apply the substitution theorem, substituting $t = as$. Now we have $D_s t = D_s as = a$, and for $t = a$, $s = 1$, and for $t = ab$, $s = b$. Thus

$$\int_a^{ab} \frac{1}{t}\,dt = \int_1^{b} \frac{1}{as}\,a\,ds = \int_1^{b} \frac{1}{s}\,ds = L(b).$$

Therefore $L(ab) = L(a) + L(b)$.

We can also see that

$$L(abc) = L(a) + L(bc) = L(a) + L(b) + L(c), \text{ etc.}$$

Therefore

$$L(a^n) = L(a) + L(a) + \ldots + L(a) = n\,L(a)$$

for n a positive integer. Moreover for n a positive integer

$$L(a) = L\left((a^{1/n})^n\right) = n\,L(a^{1/n})$$

so that $L(a^{1/n}) = \frac{1}{n} L(a)$.

It therefore follows, for n a positive rational number, that we may express r in the form $\frac{p}{q}$, with p and q positive integers so that

$$L(a^r) = L(a^{p/q}) = L(a^p)^{1/q} = \frac{1}{q} L(a^p) = \frac{1}{q} pL(a) = \frac{p}{q} L(a) = rL(a).$$

Similar results hold for negative exponents

$$L(a) + L(\frac{1}{a}) = L(a \cdot \frac{1}{a}) = L(1) = 0.$$

Therefore

$$L(\frac{1}{a}) = -L(a) \quad \text{or} \quad L(a^{-1}) = -1 \cdot L(a).$$

Thus, if r is a negative rational number (so that $-r$ is a positive rational number), we have

$$L(a^r) = L\left((a^{-r})^{-1}\right) = -1\ L(a^{-r}) = (-1)(-r)\ L(a) = r\ L(a).$$

We have shown for all rational numbers r except zero that $L(a^r) = r\ L(a)$. This formula also holds for $r = 0$ since

$$L(a^0) = L(1) = 0 = 0\ L(a).$$

Furthermore

$$L(\frac{a}{b}) = L(a) + L(\frac{1}{b}) = L(a) - L(b).$$

Two final properties of the function L follow.

Since $L(2^n) = n\ L(2)$ and $L(2)$ is a fixed positive number, we see that as n becomes large without bound, so does $L(2^n)$. Therefore $L(x) \rightarrow \infty$ as $x \rightarrow \infty$. Similarly, $L(2^{-n}) = -n\ L(2)$ so that $L(x) \rightarrow \infty$ as $x \rightarrow 0$.

We are ready to collect our results about the function L and finally to sketch its graph.

Properties of the function L:

(1) $L(x) = \int_1^x \frac{1}{t}\, dt, \quad x > 0$;

(2) L is an increasing function;

(3) $L(1) = 0$;

(4) $L(x) \rightarrow \infty$ as $x \rightarrow \infty$, and $L(x) \rightarrow \infty$ as $x \rightarrow 0$;

(5) $L'(x) = \frac{1}{x}, \quad x > 0$;

(6) $L(ab) = L(a) + L(b), \quad a,b > 0$;

(7) $L(\frac{a}{b}) = L(a) - L(b), \quad a,b > 0$;

(8) $L(a^r) = r\ L(a), \quad a > 0, \; r$ rational.

We would now be able to draw quite an accurate graph of the function L if we knew the functional value (or a close approximation of it) at one number other than the number 1. Such an approximation can be obtained as follows: First note that

$$L(1 + x) = \int_1^{1+x} \frac{1}{s}\, ds = \int_0^x \frac{1}{1 + t}\, dt$$

by the substitution theorem for integrals where we substituted $1 + t$ for s. Now, for $t > 0$ we have

$$1 - t^4 < 1 < 1 + t^5.$$

Dividing this inequality by $1 + t$ (which is positive), we have

$$1 - t + t^2 - t^3 < \frac{1}{1 + t} < 1 - t + t^2 - t^3 + t^4.$$

Therefore for $x > 0$

$$\int_0^x (1 - t + t^2 - t^3)dt < \int_0^x \frac{1}{1 + t}\,dt < \int_0^x (1 - t + t^2 - t^3 + t^4)dt.$$

Performing the integrations on left and right and noting that we have already shown that the middle integral is $L(1 + x)$, we have

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} < L(1 + x) < x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}$$

for all $x > 0$.

In particular, setting $x = 1$ yields

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} < L(2) < 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5}$$

so that

$$\left|L(2) - \left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{10}\right)\right| < \frac{1}{10}.$$

We thus see that $L(2)$ is within $\frac{1}{10}$ of $\frac{41}{60}$.

Another more natural and quite different way of computing $L(2)$ would be to approximate the integral

$$\int_1^2 \frac{1}{t}\,dt$$

by computing upper and lower sums. We use the partition

$$\Delta = \{1, \tfrac{5}{4}, \tfrac{6}{4}, \tfrac{7}{4}, 2\}.$$

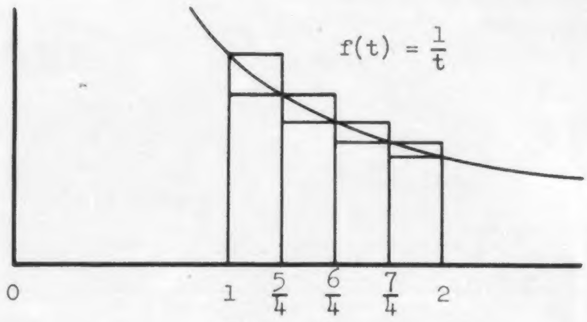


Figure 2.

Since $\frac{1}{t}$ is decreasing, we have

$$\begin{aligned}\overline{S}_{\Delta} &= M_1 \cdot \frac{1}{4} + M_2 \cdot \frac{1}{4} + M_3 \cdot \frac{1}{4} + M_4 \cdot \frac{1}{4} \\ &= 1 \cdot \frac{1}{4} + \frac{4}{5} \cdot \frac{1}{4} + \frac{4}{6} \cdot \frac{1}{4} + \frac{4}{7} \cdot \frac{1}{4} \\ &= \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}\end{aligned}$$

while

$$\begin{aligned}\underline{S}_{\Delta} &= m_1 \cdot \frac{1}{4} + m_2 \cdot \frac{1}{4} + m_3 \cdot \frac{1}{4} + m_4 \cdot \frac{1}{4} \\ &= \frac{4}{5} \cdot \frac{1}{4} + \frac{4}{6} \cdot \frac{1}{4} + \frac{4}{7} \cdot \frac{1}{4} + \frac{4}{8} \cdot \frac{1}{4} \\ &= \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}.\end{aligned}$$

Thus

$$\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} < L(2) < \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}$$

or

$$\left|L(2) - \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{3}{16}\right)\right| < \frac{1}{16}.$$

Hence $L(2)$ is within $\frac{1}{16}$ of $\frac{1066}{1580}$ or approximately $\frac{2}{3}$.

Here is the graph of L.

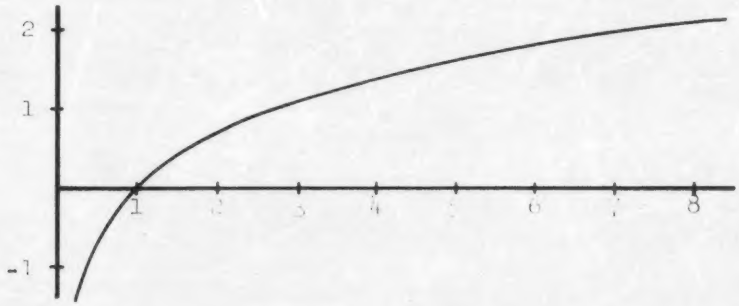


Figure 3.

In sketching this graph, we have used the following data:

$L(1) = 0$, $L'(1) = 1$, L is increasing;

$L(2) \approx \frac{2}{3}$ so that $L(4) = 2\,L(2) \approx \frac{4}{3}$, $L(8) = 3\,L(2) \approx 2$;

$L(\frac{1}{2}) = -1\,L(2) \approx -\frac{2}{3}$, $L(\frac{1}{4}) = -2\,L(2) \approx -\frac{4}{3}$.